J Syst Sci Complex (20XX) XX: 1–23

Optimal Poisson Subsampling for Softmax Regression*

YAO Yaqiong · ZOU Jiahui · WANG HaiYing

DOI:

Received: May 28 2021 / Revised: x x 20xx

Abstract Softmax regression, which is also called multinomial logistic regression, is widely used in various fields for modeling the relationship between covariates and categorical responses with multiple levels. The increasing volumes of data bring new challenges for parameter estimation in softmax regression, and the optimal subsampling method is an effective way to solve them. However, optimal subsampling with replacement requires to access all the sampling probabilities simultaneously to draw a subsample, and the resultant subsample could contain duplicate observations. In this paper, we consider Poisson subsampling for its higher estimation accuracy and applicability in the scenario that the data exceed the memory limit. We derive the asymptotic properties of the general Poisson subsampling estimator and obtain optimal subsampling probabilities by minimizing the asymptotic variance-covariance matrix under both A- and L- optimality criteria. The optimal subsampling probabilities contain unknown quantities from the full dataset, so we suggest an approximately optimal Poisson subsampling algorithm which contains two sampling steps, with the first step as a pilot phase. We demonstrate the performance of our optimal Poisson subsampling algorithm through numerical simulations and real data examples.

Keywords Multinomial Logistic Regression; Optimality Criterion; Optimal Subsampling

YAO Yaqiong
Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA.
Email: yaqiong.yao@uconn.edu
ZOU Jiahui
School of Statistics, Capital University of Economics and Business, Beijing, 100070, China.
Email: zoujiahui@amss.ac.cn
WANG HaiYing (Corresponding author)
Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA.
Email: haiying.wang@uconn.edu

° YAO Yaqiong and ZOU Jiahui contributed equally to this work.
*Wang's research was partially supported by NSF Grant CCF 2105571
◇ *This paper was recommended for publication by Editor .*

1 Introduction

The rapid development of technology makes data collection much easier than before. The growing data sizes aggravate the difficulty of data analyses because applying statistical methods directly to large datasets is sometimes computationally infeasible. A popular way to solve this issue is to use subsampling methods with non-uniform probabilities. Existing subsampling approaches mainly use sampling with replacement, which requires to access all the non-uniform probabilities at once in the sampling procedure. Compared with subsampling with replacement, Poisson subsampling manifests a great convenience because it allows to calculate subsampling probabilities and draw subsamples without loading the full data into the memory. In this paper, we focus on the softmax regression model and present an optimal sampling method based on Poisson subsampling.

Softmax regression is used to model the relationship between multi-class categorical responses and covariates. Consider a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\{y_i\}_{i=1}^N$ are categorical responses with $K+1$ possible values $c_0, c_1, c_2, ..., c_K$, and $\{\mathbf{x}_i\}_{i=1}^N$ are d dimensional covariates. A softmax regression model assumes that for each observation,

$$P(y_i = c_k|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k)}{\sum_{l=0}^K \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l)}, \tag{1}$$

where $\boldsymbol{\beta}_k, k = 0, 1, ..., K$, are d dimensional regression coefficients. For identifiability, assume that $\boldsymbol{\beta}_0 = \mathbf{0}$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, ..., \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$, a Kd dimensional vector. With this notation, the model in (1) becomes

$$\begin{aligned} P(y_i = c_0|\mathbf{x}_i) &= p_0(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + \sum_{l=1}^K \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l)}, \\ P(y_i = c_k|\mathbf{x}_i) &= p_k(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k)}{1 + \sum_{l=1}^K \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l)}, \end{aligned}$$

for $k = 1, 2, ..., K$. Denote $\delta_{i,k} = I(y_i = c_k)$ as the category indicator for the ith observation. To estimate $\boldsymbol{\beta}$, the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}_{\text{full}}$ can be obtained by maximizing

the log-likelihood function

$$\ell_f(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K}\delta_{i,k}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k - \log\left\{1+\sum_{l=1}^{K}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l)\right\}\right]. \tag{2}$$

There is no general closed-form solution to $\widehat{\boldsymbol{\beta}}_{\text{full}}$. We adopt the Newton-Raphson method to obtain $\widehat{\boldsymbol{\beta}}_{\text{full}}$ by iteratively calculating the following equation until convergence.

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + \Big\{\sum_{i=1}^{N}\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}^{(t)})\otimes(\mathbf{x}_i\mathbf{x}_t^{\mathrm{T}})\Big\}^{-1}\Big\{\sum_{i=1}^{N}\boldsymbol{s}_i(\widehat{\boldsymbol{\beta}}^{(t)})\otimes\mathbf{x}_i\Big\},$$

where $\otimes$ means the Kronecker product, $\boldsymbol{\phi}_i(\boldsymbol{\beta})$ is a $K\times K$ symmetric matrix with diagonal elements being $p_k(\mathbf{x}_i,\boldsymbol{\beta})-p_k^2(\mathbf{x}_i,\boldsymbol{\beta})$ and k_1k_2th off-diagonal elements being $-p_{k1}(\mathbf{x}_i,\boldsymbol{\beta})p_{k2}(\mathbf{x}_i,\boldsymbol{\beta})$; $\boldsymbol{s}_i(\boldsymbol{\beta})$ is a K dimensional vector with each element being $\delta_{i,k}-p_k(\mathbf{x}_i,\boldsymbol{\beta})$. It takes $O(\xi NK^2d^2+\xi K^3d^3)$ time to obtain the full data MLE, where ξ is the number of iterations for the Newton-Raphson method to converge. When $N>Kd$, the time complexity becomes $O(\xi NK^2d^2)$. For very large N, this computational cost could be high. Therefore, it is important to reduce the computational cost in estimating parameters in softmax regression for massive datasets.

One way to reduce the computational cost is through subsampling, i.e., to use a subsample of observations instead of the full dataset to perform the intended analysis. Many recent studies focus on this area. For linear regression, non-uniform subsampling was recommended in [1]. Non-uniform subsampling probabilities are often constructed by the statistical leverage scores, which can be approximated efficiently by fast randomized algorithms proposed in [2, 3]. The aforementioned algorithms were summarized as the algorithmic leveraging approach in [4]. Random projection provides another scheme to solve the estimation problem for overconstraint least squares. This approach performs randomized Hadamard transform on covariates and responses and then applies uniform subsampling method on the transformed data [5, 6]. The randomized algorithms to-date for least squares and matrix operation problems were reviewed in [7]. A deterministic approach named information-based optimal subdata selection, which selects the most informative data points characterized by the D-optimality criterion, was proposed in [8]. An extension of this method to the divide-and-conquer setting can be found in [9].

Beyond linear models, a local case control sampling method was proposed in [10] for logistic

regression with imbalanced responses. This idea was generalized in [11] to softmax regression. An optimal subsampling method under the A-optimality criterion (OSMAC) for logistic regression inspired by the idea of optimal design of experiments was developed in [12]. They proposed to use a pilot subsample to estimate the optimal subsampling probabilities, which were formatted by the asymptotic variance-covariance matrix of the subsample estimator, and then draw a second stage sample according to the estimated optimal subsampling probabilities. This method was further enhanced by using an unweighted objective function in [13]. The OSMAC was extended to include generalized linear model in [14], softmax regression in [15], Markov chain Monte Carlo in [16], quantile regression in [17], and quasi-likelihood estimation in [18]. The investigation about softmax regression in [15] focused on the subsampling with replacement. However, taking observations by optimal subsampling with replacement requires to access all subsampling probabilities at once, and the resultant subsample may contain duplicate data points. To solve these issues, we consider Poisson subsampling in this paper. Compared with subsampling with replacement, Poisson subsampling has a higher estimation accuracy and it is applicable even when the data volume is larger than the available memory.

The rest of the paper is organized as follows. Section 2 presents the general Poisson subsampling estimator and its asymptotic distribution. Section 3 shows the optimal subsampling probabilities under both A- and L- optimality criteria. Section 4 describes a practical two-step algorithm based on optimal Poisson subsampling along with its asymptotic properties. Numerical simulations and real data analyses are presented in Section 5. Section 6 gives a brief conclusion. The supplement contains all technical proofs.

2 General Poisson Subsampling Algorithm

In this section, we introduce a general Poisson subsampling algorithm in Algorithm 1 and derive the asymptotic properties of the subsampling estimator.

In this paper, we do not recommend subsampling with replacement for the following reasons. In subsampling with replacement, every draw is independent with other draws given the full data, and thus a sample could contain duplicated observations. More importantly, to implement subsampling with replacement, one has to use all subsampling probabilities at once in order to

Algorithm 1 General Poisson Subsampling Algorithm

Input: Dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and subsampling probabilities $\{n\pi_i\}_{i=1}^N$, where $\sum_{i=1}^N \pi_i = 1$ and n is the expected subsample size.
Output: Subsample estimator $\widehat{\boldsymbol{\beta}}_{\text{sub}}^P$.
Sampling:
for $i = 1$ **to** N **do**
  generate indicator variable $\nu_i \sim Bern(n\pi_i)$;
  if $\nu_i = 1$ **then**
    include $(\mathbf{x}_i, y_i)$ into sample and keep the $n\pi_i$;
  end if
end for
Denote the subsample as $\{(\mathbf{x}_i^*, y_i^*)\}_{i=1}^{n^*}$ and the corresponding subsampling probabilities as $\{n\pi_i^*\}_{i=1}^{n^*}$.
Estimation: To obtain the subsample estimator $\widehat{\boldsymbol{\beta}}_{\text{sub}}^P$, maximize the following target function

$$\ell_P^*(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{n^*}\frac{1}{n\pi_i^*}\left[\sum_{k=1}^{K}\delta_{i,k}^*\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}_i^* - \log\left\{1+\sum_{l=1}^{K}e^{\boldsymbol{\beta}_l^{\mathrm{T}}\mathbf{x}_i^*}\right\}\right].$$

The maximization is implemented through the Newton-Raphson method by iteratively applying

$$\widehat{\boldsymbol{\beta}}_{\text{sub}}^{(t+1)} = \widehat{\boldsymbol{\beta}}_{\text{sub}}^{(t)} + \left\{\sum_{i=1}^{n^*}\frac{1}{n\pi_i^*}\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{(t)})\otimes\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}}\right\}^{-1}\left\{\sum_{i=1}^{n^*}\frac{1}{n\pi_i^*}\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{(t)})\otimes\mathbf{x}_i^*\right\}$$

until convergence.

draw a subsample. This would be difficult when the size of the full data exceeds the memory's limit. To solve the aforementioned limitations, we consider Poisson subsampling, which decides if one observation is included in the sample or not by conducting a Bernoulli trail. Therefore, Poisson subsampling ensures no repeated data points in the sample. Moreover, for Poisson subsampling, it is possible to determine if one observation should be included or not in the subsample without accessing other data points. One limitation of Poisson subsampling is that the subsample size is random. However, we can control the expectation of the random subsample size, which we call the expected subsample size. In Algorithm 1, the expected subsample size is denoted as n and the actual subsample size is denoted as n^*.

Algorithm 1 becomes the uniform Poisson subsampling by choosing $\pi_i = \frac{1}{N}, i = 1, 2, ..., N$, which means all observations are treated equally. However, in order to obtain better approximation of $\widehat{\boldsymbol{\beta}}_{\text{full}}$, we want the subsample to include more informative observations, so we prefer non-uniform subsampling probabilities.

We derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{P}$, for which we need the following assumptions.

Assumption 1 The parameter space Θ of $\boldsymbol{\beta}$ is a compact set.

Assumption 2 As N goes to ∞, the negative Hessian matrix $\mathbf{M}_N = N^{-1}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})$ goes to a positive-definite matrix in probability.

Assumption 3 $N^{-1}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3 = O_P(1)$.

Assumption 4 For $k = 0$ and 4, $N^{-2}\sum_{i=1}^{N}\pi_i^{-1}\|\mathbf{x}_i\|^k = O_P(1)$; and there exists some $\delta > 0$ such that $N^{-(2+\delta)}\sum_{i=1}^{N}\pi_i^{-1-\delta}\|\mathbf{x}_i\|^{2+\delta} = O_P(1)$.

Assumption 2 ensures that the Hessian matrix $\mathbf{M}_N$ is invertible as $N \to \infty$. Assumption 3 tells that the third moment of covariates is bounded in probability. Assumption 4 constrains the relationship between covariates and subsampling probabilities.

Theorem 2.1 *Under Assumptions 1, 2, 3 and 4, given the full data $\mathcal{D}_N$, as $N \to \infty$ and $n \to \infty$, the conditional distribution of $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{P} - \widehat{\boldsymbol{\beta}}_{\text{full}}$ is asymptotically normal, namely,*

$$\sqrt{n}\mathbf{V}_G^{-1/2}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{P} - \widehat{\boldsymbol{\beta}}_{\text{full}}) \to \mathbb{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

in distribution, where $\mathbf{V}_G = \mathbf{M}_N^{-1}\mathbf{V}_{PG}\mathbf{M}_N^{-1}$,

$$\mathbf{M}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}), \tag{4}$$

$$\mathbf{V}_{PG} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{(1-n\pi_i)\{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\}}{\pi_i}, \tag{5}$$

and $\psi_i(\boldsymbol{\beta})$ is a $K \times K$ matrix with $\psi_{i,k_1,k_2}(\boldsymbol{\beta}) = \{\delta_{i,k_1} - p_{k_1}(\mathbf{x}_i,\boldsymbol{\beta})\}\{\delta_{i,k_2} - p_{k_2}(\mathbf{x}_i,\boldsymbol{\beta})\}$.

3 Optimal Subsampling Probabilities

To improve the estimation efficiency, we formulate the optimal subsampling probabilities by minimizing the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{P} - \widehat{\boldsymbol{\beta}}_{\text{full}}$, which is $n^{-1}\mathbf{V}_G$. Since $\mathbf{V}_G$ is a matrix, we adopt the A-optimality criterion in optimal design to minimize the trace of the variance-covariance matrix, which is $\text{tr}(n^{-1}\mathbf{V}_G)$ in our case. This quantity is often called the asymptotic mean squared error (MSE). In addition to the A-optimality, we consider the

L-optimality, which is to minimize the trace of the variance-covariance matrix of a linearly transformed estimator. Here we decide to minimize $\mathrm{tr}(\mathbf{M}_N\mathbf{V}_G\mathbf{M}_N) = \mathrm{tr}(\mathbf{V}_{PG})$ due to the computational benefit as seen later. In the following of the paper, we use $^{\mathrm{optA}}$ to represent quantities associated with the A-optimality criterion and use $^{\mathrm{optL}}$ to represent quantities associated with the L-optimality criterion. Before presenting the optimal subsampling probabilities, we define the following two notations to facilitate the presentation.

$$t_i^{\mathrm{optA}} = \|M_N^{-1}\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\mathrm{full}}) \otimes \mathbf{x}_i\}\|, \quad i = 1, 2, ..., N; \tag{6}$$

$$t_i^{\mathrm{optL}} = \|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\mathrm{full}})\|\|\mathbf{x}_i\|, \quad i = 1, 2, ..., N; \tag{7}$$

furthermore, define $t_{N+1}^{\mathrm{optA}} = +\infty$ and $t_{N+1}^{\mathrm{optL}} = +\infty$.

Theorem 3.1 *Denote the order statistic of* $\{t_i\}_{i=1}^{N+1}$ *as* $t_{(1)}, t_{(2)}, ..., t_{(N+1)}$*, which are arranged in an increasing way. Under selected optimality criterion, the optimal subsampling probabilities are*

$$\pi_i^{\mathrm{opt}} = \frac{t_i \wedge H}{\sum_{j=1}^{N}(t_j \wedge H)}, \quad i = 1, 2, ..., N, \tag{8}$$

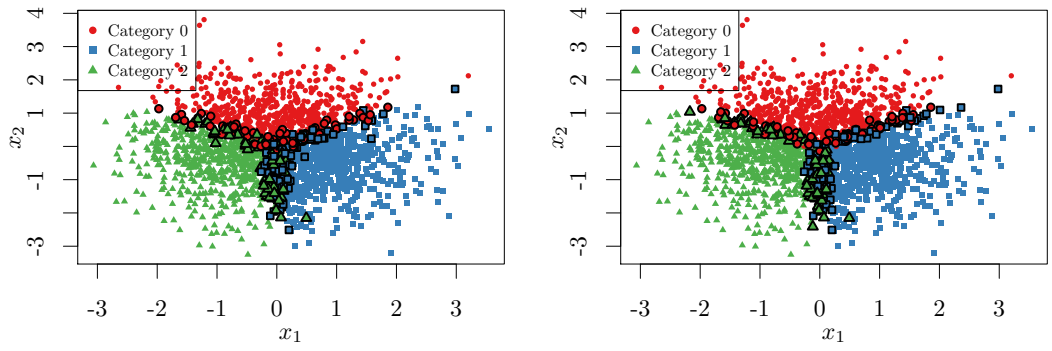
where $t_i \wedge H = \min(t_i, H)$*,*

$$H = \frac{\sum_{i=1}^{N-g} t_{(i)}}{n-g}, \tag{9}$$

and g *is an integer satisfying*

$$\frac{(n-g)t_{(N-g)}}{\sum_{i=1}^{N-g} t_{(i)}} < 1, \quad \frac{(n-g+1)t_{(N-g+1)}}{\sum_{i=1}^{N-g+1} t_{(i)}} \geq 1. \tag{10}$$

In Theorem 3.1, H works as a threshold to make sure that none of $\{n\pi_i^{\mathrm{opt}}\}_{i=1}^{N}$ is larger than 1. For the ith observation with $t_i > H$, its optimal subsampling probability $n\pi_i$ equals 1, which means that this observation will be included in the subsample for sure. Moreover, the expressions of t_i^{optA} and t_i^{optL} show that L-optimality is more computationally efficient than A-optimality because the time complexity of calculating $\{t_i^{\mathrm{optA}}\}_{i=1}^{N}$ is $O(NK^2d^2)$, whereas the time complexity of calculating $\{t_i^{\mathrm{optL}}\}_{i=1}^{N}$ is $O(NKd)$.



(a) Sampling observations by π_i^{optA} (b) Sampling observations by π_i^{optL}

Figure 1 Illustrate the distribution of chosen samples by a simulated dataset where covariates are generated from $\mathbb{N}_2(\mathbf{0}, \mathbf{I}_2)$ and the true coefficient is $\boldsymbol{\beta} = (5\sqrt{3}, -15, -5\sqrt{3}, -15)^{\mathrm{T}}$ when $N = 2000$ and $n = 200$ under both A-optimality criterion and L-optimality criterion. Observations from the category 0 are represented by red filled circle. Observations from the category 1 are represented by blue filled square. Observations from the category 2 are represented by green filled triangle point-up. The drawn samples are indicated by black boundary.

We investigate the distribution of the sample drawn by the optimal subsampling probabilities by a simulated dataset in Figure 1. It shows that the observations in the transition area of two categories are more likely to be chosen under both A-optimality criterion and L-optimality criterion. In principle, this pattern can be learned from (7) in a sense that higher optimal subsampling probability for one observation comes from larger value of $|\delta_{i,k} - p_{i,k}(\boldsymbol{\beta})|, k = 1, 2, ..., K$. If one observation with higher probability falling in category k falls into another category, then it has greater opportunity to be selected. Similarly, if one observation comes from category k where the observation has lower probability to be fallen in, then it is more likely to be sampled. This means observations, which are more likely to be miss-classified, have more chance to be drawn into the sample.

4 Approximately Optimal Poisson Subsampling Algorithm

The optimal subsampling probabilities are used to draw subsamples containing more informative observations so that we can obtain better approximations of $\widehat{\boldsymbol{\beta}}_{\text{full}}$. However, the

computation of $\{\pi_i^{\text{opt}}\}_{i=1}^N$ in Theorem 3.1 involves $\widehat{\boldsymbol{\beta}}_{\text{full}}$, which is the unknown quantity to be approximated by the subsample. To deal with this problem, we propose an approximately optimal Poisson subsampling algorithm which uses a pilot sample estimator to substitute the $\widehat{\boldsymbol{\beta}}_{\text{full}}$ in Theorem 3.1. We draw the pilot subsample according to the proportion-based subsampling probabilities $\{n_0\pi_i^{\text{prop}}\}_{i=1}^N$, where $\pi_i^{\text{prop}} = \sum_{k=0}^K \frac{\delta_{i,k}}{(K+1)m_k}$, m_k is the number of responses in the kth category, and n_0 is the expected sample size of the pilot subsample. We could also use the uniform subsampling probabilities. However, for imbalanced datasets, the probability that some categories contribute no observation to the subsample can be high.

Let the pilot sample of actual size n_0^* be $\{\mathbf{x}_i^{*0}, y_i^{*0}\}_{i=1}^{n_0^*}$ and the corresponding subsampling probabilities be $\{n_0\pi_i^{*0}\}_{i=1}^{n_0^*}$. After obtaining the pilot estimator $\widehat{\boldsymbol{\beta}}_P^0$ from the pilot sample, the optimal subsampling probabilities can be approximated by

$$n\widetilde{\pi}_i^{\text{opt}} = \frac{n(\widehat{t}_i \wedge \widehat{H})}{\frac{n_0^*}{n_0^*-d\times K}\sum_{i=1}^{n_0^*}\frac{\widehat{t}_{0i}\wedge\widehat{H}}{n_0^*\pi_i^{*0}}} \wedge 1, \tag{11}$$

where $\{\widehat{t}_{0i}\}_{i=1}^{n_0^*}$ is either $\{\widehat{t}_{0i}^{\text{optA}}\}_{i=1}^{n_0^*}$ or $\{\widehat{t}_{0i}^{\text{optL}}\}_{i=1}^{n_0^*}$ with $\widehat{t}_{0i}^{\text{optA}} = \|(\widehat{\mathbf{M}}_N^0)^{-1}\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_P^0)\otimes\mathbf{x}_i^{*0}\}\|$ or $\widehat{t}_{0i}^{\text{optL}} = \|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_P^0)\|\|\mathbf{x}_i^{*0}\|$, respectively;

$$\widehat{\mathbf{M}}_N^0 = \frac{1}{N}\sum_{i=1}^{n_0^*}\frac{\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_P^0)\otimes\{\mathbf{x}_i^{*0}(\mathbf{x}_i^{*0})^{\text{T}}\}}{n_0^*\pi_i^{*0}}; \tag{12}$$

$\widehat{H}$ is the $(1-\frac{n}{2N})$th quantile of $\{\widehat{t}_{0i}\}_{i=1}^{n_0^*}$; $\frac{n_0^*}{n_0^*-d\times K}$ is a finite-sample term to correct the degrees of freedom; and $\widehat{t}_i$ is $\widehat{t}_i^{\text{optA}}$ or $\widehat{t}_i^{\text{optL}}$ according to the selected optimality criterion with the following expressions

$$\widehat{t}_i^{\text{optA}} = \|(\widehat{\mathbf{M}}_N^0)^{-1}\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_P^0)\otimes\mathbf{x}_i\}\| \quad \text{and} \quad \widehat{t}_i^{\text{optL}} = \|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_P^0)\|\|\mathbf{x}_i\|.$$

The approximately optimal Poisson subsampling algorithm is presented in Algorithm 2.

Remark 4.1 To reduce the computational burden when approximating the optimal subsampling probabilities in (8), we use $\widehat{\mathbf{M}}_N^0$ to estimate $\mathbf{M}_N$ and use the $(1-\frac{n}{2N})$th quantile

Algorithm 2 Approximately Optimal Poisson Subsampling Algorithm

First Stage Sampling: Run Algorithm 1 with expected sample size n_0 and subsampling probabilities $\{n_0\pi_i^{\text{prop}}\}_{i=1}^N$ to obtain the first stage sample $\{\mathbf{x}_i^{*0}, y_i^{*0}\}_{i=1}^{n_0^*}$ and the corresponding subsampling probabilities $\{n_0\pi_i^{*0}\}_{i=1}^{n_0^*}$, and use them to obtain the coefficient estimator $\widehat{\boldsymbol{\beta}}_P^0$, where n_0^* is the actual subsample size.

Second Stage Sampling: Run Algorithm 1 with expected sample size n and the approximated optimal subsampling probabilities $\{n\widetilde{\pi}_i^{\text{opt}}\}_{i=1}^N$ in (11). Obtain the second stage subsample $\{\mathbf{x}_i^{*1}, y_i^{*1}\}_{i=1}^{n^*}$ and the corresponding subsampling probabilities $\{n\pi_i^{*1}\}_{i=1}^{n^*}$, where n^* is the actual subsample size. Use the second stage subsample to calculate the coefficient estimator $\widehat{\boldsymbol{\beta}}_P^1$.

Combining: Obtain the final estimator $\widehat{\boldsymbol{\beta}}_P^{ada}$ by combining $\widehat{\boldsymbol{\beta}}_P^0$ and $\widehat{\boldsymbol{\beta}}_P^1$, through

$$\widehat{\boldsymbol{\beta}}_P^{ada} = (n_0^*\widehat{\mathbf{M}}_N^0 + n^*\widehat{\mathbf{M}}_N^1)^{-1}(n_0^*\widehat{\mathbf{M}}_N^0\widehat{\boldsymbol{\beta}}_P^0 + n^*\widehat{\mathbf{M}}_N^1\widehat{\boldsymbol{\beta}}_P^1),$$

where $\widehat{\mathbf{M}}_N^0$ is defined in (12), and

$$\widehat{\mathbf{M}}_N^1 = \frac{1}{N}\sum_{i=1}^{n^*}\frac{\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_P^1)\otimes\{\mathbf{x}_i^{*1}(\mathbf{x}_i^{*1})^{\mathrm{T}}\}}{n^*\pi_i^{*1}}.$$

of $\{\widehat{t}_i\}_{i=1}^{n_0^*}$ to estimate H, instead of directly substituting $\widehat{\boldsymbol{\beta}}_{\text{full}}$ with $\widehat{\boldsymbol{\beta}}_P^0$ in (4) and (9). The numerical results in Section 5 show that the performance of the resulting estimator is similar to that of the estimator obtained by substituting $\widehat{\boldsymbol{\beta}}_{\text{full}}$ with $\widehat{\boldsymbol{\beta}}_P^0$ in (4) and (9). In addition, an even simpler approach is to treat $\widehat{H}$ as infinity, and we will evaluate the performance of this choice in Section 5 as well.

Remark 4.2 In Algorithm 2, the final estimator is obtained by combining the two subsample estimators in the two stages. We could also obtain the final estimator by combining the subsamples from the two stages. The performances of these combining methods are similar in terms of estimation efficiency when n_0 and n are large. However, combining estimators waives the need to apply Newton-Raphson calculations on the first stage subsample twice. When n_0 and n are small, combining samples could be more computationally stable since the Newton-Raphson method is applied to a larger sample.

To derive the asymptotic property of $\widehat{\boldsymbol{\beta}}_P^{ada}$, we need the following assumption.

Assumption 5 The covariates $\{\mathbf{x}_i\}_{i=1}^N$ are independent and identically distributed random variables. $\mathbb{E}(\|\mathbf{x}_i\|^{-2}) < \infty$ and there exists a constant $c > 0$ such that $\mathbb{E}(e^{a\|\mathbf{x}_i\|}) \leq \exp(c^2a^2/2)$ for all $a \in \mathbf{R}$.

This assumption constrains the distribution of the covariates. If we include intercept in the model, the condition $\mathbb{E}(\|\mathbf{x}_i\|^{-2}) < \infty$ is always satisfied because $\|\mathbf{x}_i\| \geq 1$ almost surely in this scenario.

Theorem 4.3 *Under Assumptions 1, 2 and 5, if $n = o(N/\ln N)$ and $n_0 = o(n^{1/2})$, as $n_0 \to \infty$, $n \to \infty, N \to \infty$, conditioned on $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}_P^0$,*

$$\sqrt{n}\mathbf{V}^{-1/2}\left(\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}}\right) \to \mathbb{N}(\mathbf{0}, \mathbf{I}) \tag{13}$$

in distribution, where $\widehat{\boldsymbol{\beta}}_P^{ada}$ is obtained with $H = +\infty$ under L-optimality criterion, and $\mathbf{V} = \mathbf{M}_N^{-1}\mathbf{V}_P\mathbf{M}_N^{-1}$, with $\mathbf{V}_P$ having the expression of

$$\mathbf{V}_P = \frac{1}{N^2}\left[\sum_{i=1}^{N}\frac{\{1 - n\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\}\{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\}}{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_i\|}\right]\left[\sum_{j=1}^{N}\|\mathbf{s}_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_j\|\right].$$

Combining with Theorem 3.1, Theorem 4.3 tells us that, theoretically, the approximately optimal Poisson subsampling should have a better estimation efficiency than uniform Poisson subsampling since $\text{tr}(\mathbf{V}_P)$ is smaller than $\text{tr}(\mathbf{V}_{PG})$ for $\mathbf{V}_{PG}$ with $\pi_i = 1/N$ in Theorem 2.1. Moreover, [15] presented an optimal subsampling with replacement algorithm without deriving the asymptotic distribution of the final estimator. To compare our results based on optimal Poisson subsampling with that of [15], we derive the asymptotic distribution for the estimator from their algorithm. Since the theoretical properties of [15]'s algorithm is not the main focus of our paper, we put the results in the supplement. From Theorem A.2.1 of Section A.2 in the supplement, the asymptotic variance-covariance matrix (scaled by n) of the estimator in [15] is $\mathbf{V}_S = \mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}$, where

$$\mathbf{V}_{Nc} = \frac{1}{N^2}\left[\sum_{i=1}^{N}\frac{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_i\|}\right]\left[\sum_{j=1}^{N}\|\mathbf{s}_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_j\|\right].$$

Compared with $\mathbf{V}_{Nc}$, $\mathbf{V}_P$ has one more subtraction term. Since $\mathbf{V}_S > \mathbf{V}$ in the Loewner ordering, the approximately optimal Poisson subsampling algorithm is more efficient than the optimal subsampling with replacement algorithm.

The asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_P^{ada}$ is $n^{-1}\mathbf{V}$ in Theorem 4.3, which depends

on $\widehat{\boldsymbol{\beta}}_{\text{full}}$. So we can insert $\widehat{\boldsymbol{\beta}}_P^0$ and $\widehat{\boldsymbol{\beta}}_P^1$ in order to approximate the variance-covariance matrix. To accelerate the computation, we recommend using the subsamples from the two stages instead of using the full data to approximate the variance-covariance matrix. Specifically, the variance-covariance matrix can be approximated by

$$\breve{\mathbf{V}} = (\widehat{\mathbf{M}}_N^0 + \widehat{\mathbf{M}}_N^1)^{-1}(\widehat{\mathbf{V}}_P^0 + \widehat{\mathbf{V}}_P^1)(\widehat{\mathbf{M}}_N^0 + \widehat{\mathbf{M}}_N^1)^{-1}, \tag{14}$$

where

$$\widehat{\mathbf{V}}_P^0 = \frac{1}{N^2}\sum_{i=1}^{n_0^*} \frac{(1 - n_0\pi_i^{*0})\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_P^0) \otimes \mathbf{x}_i^{*0}(\mathbf{x}_i^{*0})^{\mathrm{T}}}{(\pi_i^{*0})^2},$$
$$\widehat{\mathbf{V}}_P^1 = \frac{1}{N^2}\sum_{i=1}^{n^*} \frac{(1 - n\pi_i^{*1})\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_P^1) \otimes \mathbf{x}_i^{*1}(\mathbf{x}_i^{*1})^{\mathrm{T}}}{(\pi_i^{*1})^2}.$$

In [15], the authors did not mention how to estimate the variance-covariance matrix of their estimator, say $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}$, based on optimal subsampling with replacement. To compare the performances of Poisson subsampling and subsampling with replacement, we propose to approximate the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}$ by

$$\breve{\mathbf{V}}_{\text{sub}} = \breve{\mathbf{M}}_{N,\text{sub}}^{-1}\breve{\mathbf{V}}_{Nc}\breve{\mathbf{M}}_{N,\text{sub}}^{-1}, \tag{15}$$

where

$$\breve{\mathbf{M}}_{N,\text{sub}} = \frac{1}{N}\sum_{i=1}^{n_0+n} \frac{\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}) \otimes \{\mathbf{x}_i^{*\text{sub}}(\mathbf{x}_i^{*\text{sub}})\}^{\mathrm{T}}}{\pi_i^{*\text{sub}}}, \tag{16}$$

$$\breve{\mathbf{V}}_{Nc} = \frac{1}{N^2}\sum_{i=1}^{n_0+n} \frac{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}) \otimes \{\mathbf{x}_i^{*\text{sub}}(\mathbf{x}_i^{*\text{sub}})\}^{\mathrm{T}}}{(\pi_i^{*\text{sub}})^2}, \tag{17}$$

and $\mathbf{x}_i^{*\text{sub}}$ and $\pi_i^{*\text{sub}}$ are the subsample observations and the corresponding subsampling probabilities obtained by the method proposed in [15]. The performance of $\breve{\mathbf{V}}$ and $\breve{\mathbf{V}}_{\text{sub}}$ will be evaluated in Section 5.1.1.

5 Numerical Results

In this section, we evaluate the performance of Algorithm 2 and its two variants discussed in Remark 4.1 by using simulated and real datasets. We also apply the approximately optimal subsampling with replacement algorithm proposed in [15] for comparison.

5.1 Simulations

We set the full data size $N = 100000$ and assume that the responses $\{y_i\}_{i=1}^N$ consist of three distinct outcomes $0, 1, 2$. The dimension of the covariate is $d = 3$ and the true value of the parameter is $\boldsymbol{\beta} = (1, 1, 1, 2, 2, 2)^{\mathrm{T}}$. We consider the following four cases of covariate structures, where $\boldsymbol{\Sigma}$ is a 3×3 matrix with diagonal elements being 1 and off-diagonal elements being 0.5.

Case 1. $\mathbf{x}_i \sim \mathbb{N}_3(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, 2, ..., N$. From this structure, the generated responses $\{y_i\}_{i=1}^N$ have roughly 42% of observations in each of the first and the third categories, and around 16% of observations are in the second category.

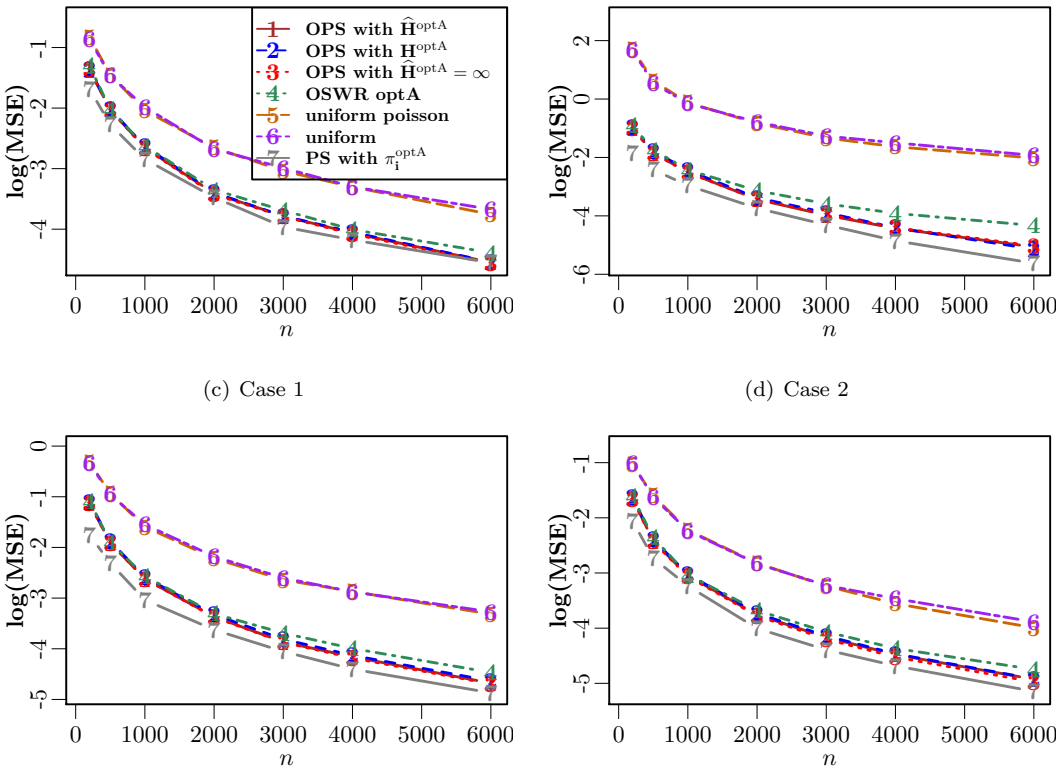
Case 2. $\mathbf{x}_i \sim \mathbb{N}_3(\mathbf{1.5}, \boldsymbol{\Sigma}), i = 1, 2, ..., N$. The location shift of the covariate distribution results in imbalanced responses. About 90% observations falls into the third category.

Case 3. $\mathbf{x}_i \sim 0.5\mathbb{N}_3(\mathbf{1}, \boldsymbol{\Sigma}) + 0.5\mathbb{N}_3(\mathbf{-1}, \boldsymbol{\Sigma}), i = 1, 2, ..., N$. This is a mixture normal distribution. Around 45% data falls into the first category; 10% data falls into the second category and 45% data falls into the third category.

Case 4. $\mathbf{x}_i \sim t_3(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, 2, ..., N$. The t_3 distribution has a heavier tail than normal distribution. Around 42.5% data falls into the first category; 15% data falls into the second category and 42.5% data falls into the third category.

We are going to assess the estimation efficiency and computation efficiency of the proposed algorithms based on these four datasets.

5.1.1 Estimation Efficiency



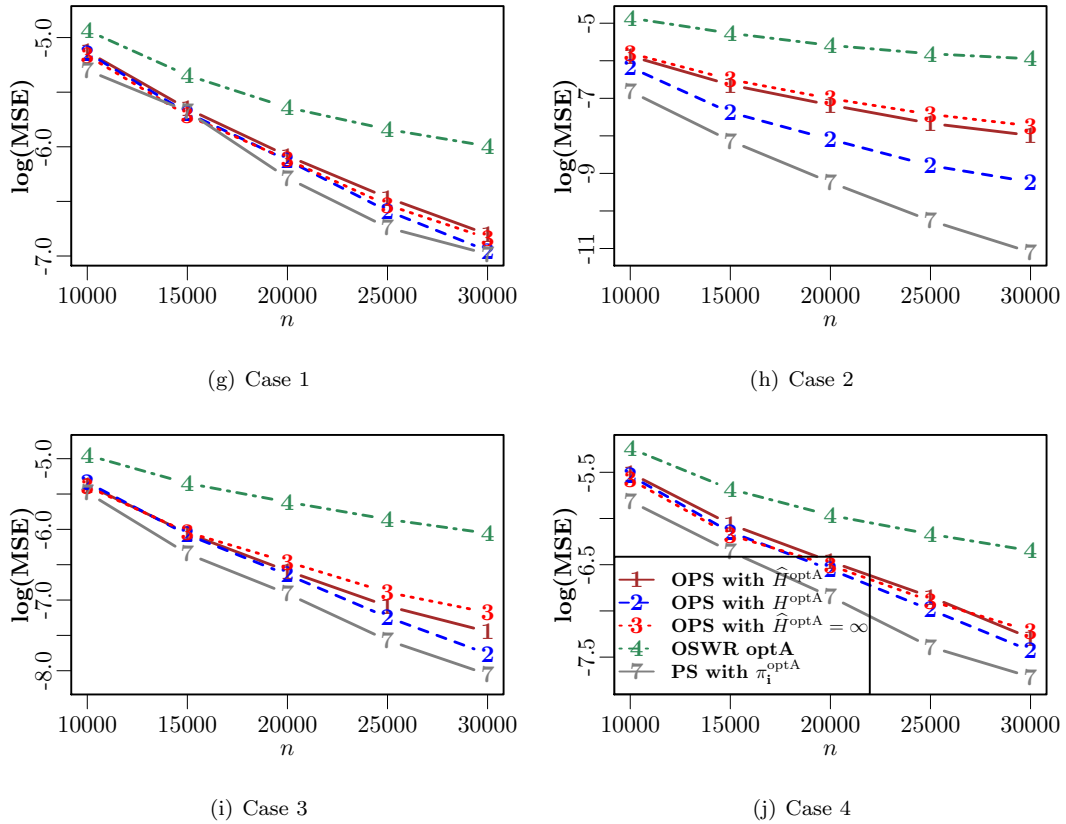
(c) Case 1 (d) Case 2

(e) Case 3 (f) Case 4

Figure 2 Empirical MSEs among different n when $n_0 = 200$ is fixed using different methods. OPS with $\widehat{H}^{\text{optA}}$ stands for Algorithm 2 under the A-optimality criterion; OPS with H^{optA} stands for Algorithm 2 under the A-optimality criterion by directly substituting $\widehat{\boldsymbol{\beta}}_{\text{full}}$ with $\widehat{\boldsymbol{\beta}}_P^0$ in (4) and (9); OPS with $\widehat{H}^{\text{optA}} = \infty$ stands for Algorithm 2 under the A-optimality criterion with taking $\widehat{H}^{\text{optA}} = \infty$; OSWR optA stands for approximately optimal subsampling with replacement algorithm under the A-optimality criterion. PS with π_i^{optA} stands for Algorithm 1 by using the optimal subsampling probabilities under A-optimality criterion. All one stage subsampling algorithms take $(n_0 + n)$ as the subsample size or the expected subsample size.

The estimation efficiency is evaluated by empirical MSE, defined as $\text{MSE} = S^{-1}\sum_{s=1}^{S}\|\widetilde{\boldsymbol{\beta}}_s - \widehat{\boldsymbol{\beta}}_{\text{full}}\|^2$, where S is the number of replicates and $\widetilde{\boldsymbol{\beta}}_s$ is the estimate in the sth replication. Figure 2 shows that the approximately optimal subsampling algorithms have an obvious benefit in

estimation accuracy compared with the uniform probabilities based algorithms. Among these four approximately optimal subsampling algorithms, as n goes large, the three algorithms based on Poisson subsampling show a slight advantage to the algorithm based on subsampling with replacement. The plot for Case 4 shows that even when Assumption 5 is not satisfied, the approximately optimal subsampling algorithms still perform well. Algorithm 1 with optimal subsampling probabilities acts as a reference and does not have practical utility because the calculation of π_i^{optA} involves $\widehat{\boldsymbol{\beta}}_{\text{full}}$. The four approximately optimal subsampling algorithms perform closely to Algorithm 1 with optimal subsampling probabilities showing that using pilot sample estimator to substitute $\widehat{\boldsymbol{\beta}}_{\text{full}}$ in Theorem 3.1 costs little in estimation accuracy.



(g) Case 1 (h) Case 2

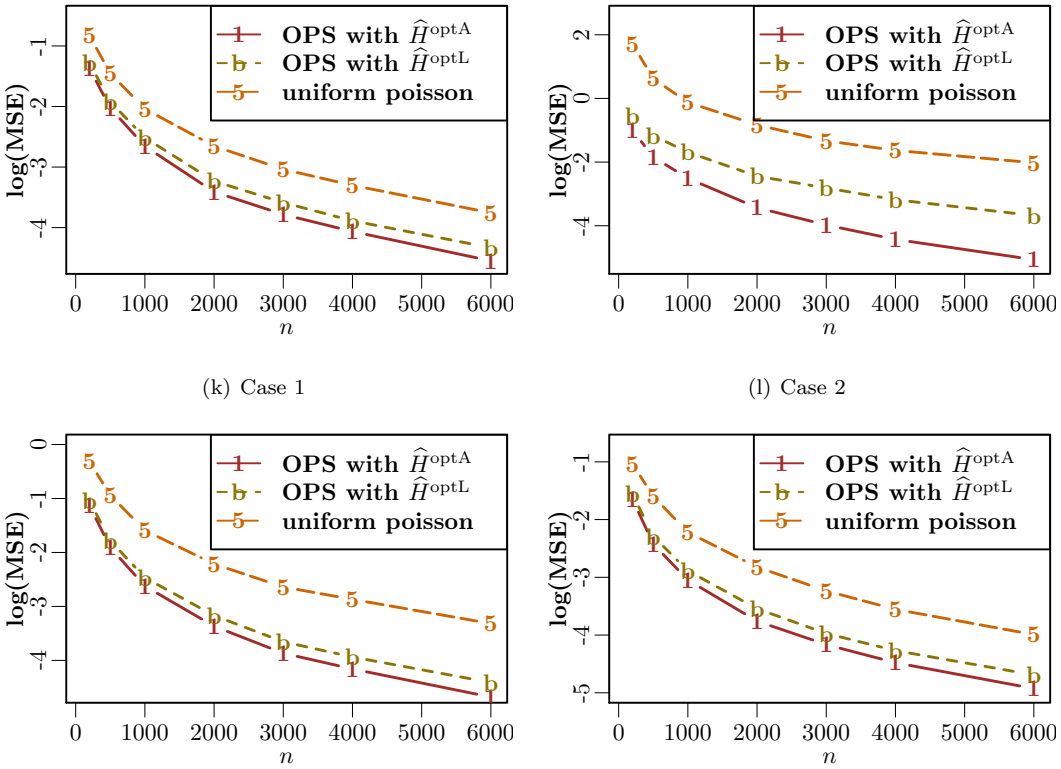
(i) Case 3 (j) Case 4

Figure 3 Empirical MSEs of different methods when $n_0 = 200$ is fixed and n is large.

Figure 3 further compares the four approximately optimal subsampling algorithms when n is relatively large. The three algorithms based on Poisson subsampling dominate the algorithm based on subsampling with replacement, and this superiority becomes more significant as n

becomes larger. Among these three algorithms based on Poisson subsampling, overall the one with H^{optA} preforms best, which is reasonable because it uses the full data instead of the pilot subsample to estimate H and therefore should result in better approximations to the optimal subsampling probabilities.

Figure 4 evaluates the estimation performance of Algorithm 2 under different optimality criteria, and shows that the Algorithm 2 under the A-optimality criterion is more efficient than that under the L-optimality criterion in terms of MSE. This is reasonable because $\{\pi_i^{\text{optA}}\}_{i=1}^N$ aims at minimizing the asymptotic MSE.



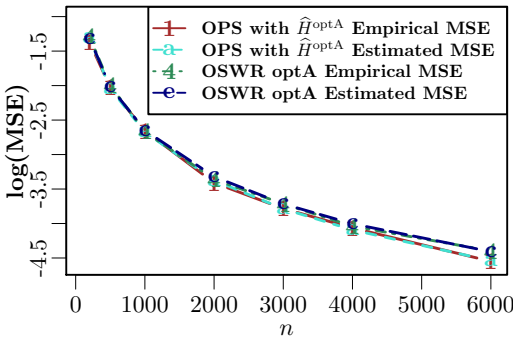
(k) Case 1 (l) Case 2

(m) Case 3 (n) Case 4

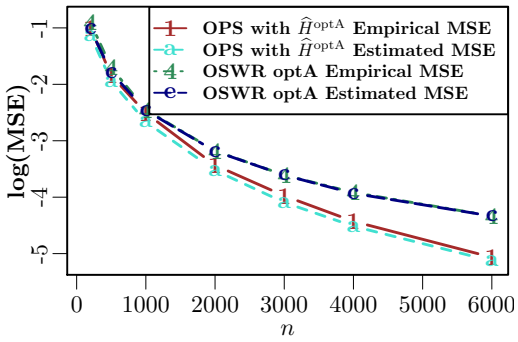
Figure 4 Empirical MSEs among different n when $n_0 = 200$ is fixed using different methods and different optimality criteria. OPS with $\widehat{H}^{\text{optA}}$ means Algorithm 2 under A-optimality criterion; OPS with $\widehat{H}^{\text{optL}}$ means Algorithm 2 under L-optimality criterion.

To demonstrate the performance of the formulas (14) and (15) in estimating the variance-covariance matrices, we compare the empirical MSEs with the estimated MSEs for all four cases.

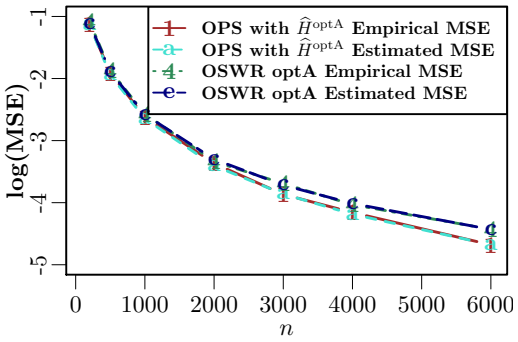
Here an estimated MSE is the trace of the estimated variance-covariance matrix. Figure 5 shows that the proposed formula in (14) works well for Algorithm 2, and the proposed formula in (15) works well for the approximately optimal algorithm based on subsampling with replacement. The results for using the L-optimality criterion are omitted due to similarity.



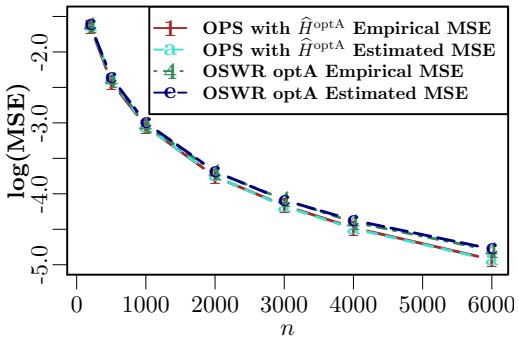
(o) Case 1



(p) Case 2



(q) Case 3



(r) Case 4

Figure 5 Empirical MSEs and estimated MSEs for two different subsampling methods with a fixed $n_0 = 200$ and different n.

5.1.2 Computation Efficiency

Besides estimation efficiency, we also investigate computational efficiency by comparing the CPU seconds that each algorithm uses in Case 1 on a MacBook Pro equipped with a 2.5 GHz Intel Core i7 processor and 16 GB memory using R [19]. Table 1 indicates that the approximately optimal subsampling algorithms under the L-optimality criterion are always faster than that under the A-optimality criterion. This is as expected because compared with π_i^{optA}, π_i^{optL} has a simpler expression and its calculation does not involve matrix multiplication. Among all approximately optimal subsampling algorithms, Poisson subsampling algorithm with

H^{optA} takes the longest time which is reasonable because it approximates M_N and H using the full data instead of the pilot subsample. The uniform subsampling algorithms (one based on Poisson subsampling and the other based on sampling with replacement) take the least time because there is no overhead time to calculate the subsampling probabilities. Clearly, using full data to compute $\widehat{\boldsymbol{\beta}}_{\text{full}}$ directly is the most time-consuming method.

Table 1 CPU seconds of different algorithms for Case 1 with a fixed $n_0 = 200$ and different n. The full data sample size is $N = 10^5$ with three categories for the responses and covariate dimension $d = 3$. The times are the total times calculated from 1000 implementations of each algorithms.

Method	n		
	200	1000	3000
OPS with $\widehat{H}^{\text{optA}}$	121.78	123.75	132.61
OPS with $\widehat{H}^{\text{optL}}$	104.52	106.83	115.17
OPS with $\widehat{H}^{\text{optA}} = \infty$	120.97	122.85	131.46
OPS with $\widehat{H}^{\text{optL}} = \infty$	103.74	106.13	113.94
OPS with H^{optA}	148.58	149.54	158.80
OPS with H^{optL}	110.05	112.32	120.74
OSWR optA	120.82	123.00	125.57
OSWR optL	97.80	99.63	109.64
Uniform Poisson	8.95	10.88	17.86
Uniform	6.29	8.19	14.80

Next, we compare the computational efficiency among different algorithms for larger data volumes by increasing the number of categories of the response variable and the dimension of the covariates. We investigate how the computational cost changes as the full data sample size N goes large. Suppose that the response variable has 6 categories, meaning $K = 5$, and the dimension of covariates is $d = 10$. Let the true value of $\boldsymbol{\beta}$ be $(\mathbf{1}_{10}^{\mathrm{T}}, 2\mathbf{1}_{10}^{\mathrm{T}}, 3\mathbf{1}_{10}^{\mathrm{T}}, 4\mathbf{1}_{10}^{\mathrm{T}}, 5\mathbf{1}_{10}^{\mathrm{T}})^{\mathrm{T}}$,

where $\mathbf{1}_{10}$ is a 10 dimensional vector with each entry being 1. We generate the covariates from $\mathbb{N}_{10}(\mathbf{0}, \mathbf{\Sigma}_{10})$ with $\mathbf{\Sigma}_{10} = 0.5\mathbf{I}_{10} + 0.5\mathbf{J}_{10}$, where $\mathbf{J}_{10}$ is a 10×10 matrix with each entry being 1. Table 2 shows that all approximately optimal subsampling algorithms demonstrate a superior computational efficiency to the full data computation. We also implement the stochastic gradient descent (SGD) method for comparison. SGD is nearly twice faster than full data computation and much slower than the approximately optimal subsampling algorithms.

Table 2 CPU seconds for different algorithms with $n_0 = 1000$, $n = 2000$, and different N. The response variable contains 6 categories and covariates are generated from $\mathbb{N}_{10}(\mathbf{0}, \mathbf{\Sigma}_{10})$. The times are the total times calculated from 20 implementations of each algorithms. SGD stands for the stochastic gradient descent with a learning rate 0.001.

Method	N		
	10^5	5×10^5	10^6
OPS with $\widehat{H}^{\text{optA}}$	15.99	65.70	126.77
OPS with $\widehat{H}^{\text{optL}}$	5.84	15.89	28.53
OPS with $\widehat{H}^{\text{optA}} = \infty$	15.85	66.15	126.64
OPS with $\widehat{H}^{\text{optL}} = \infty$	5.78	15.73	28.70
OPS with H^{optA}	23.23	101.12	196.85
OPS with H^{optL}	5.89	16.42	29.84
OSWR optA	16.75	65.72	127.49
OSWR optL	6.53	16.17	28.42
Uniform Poisson	3.39	3.92	4.56
Uniform	3.25	3.46	3.88
SGD	65.82	331.33	657.44
Full data	107.61	567.26	1146.82

5.2 Real Data Analysis

5.2.1 Cover Type Data

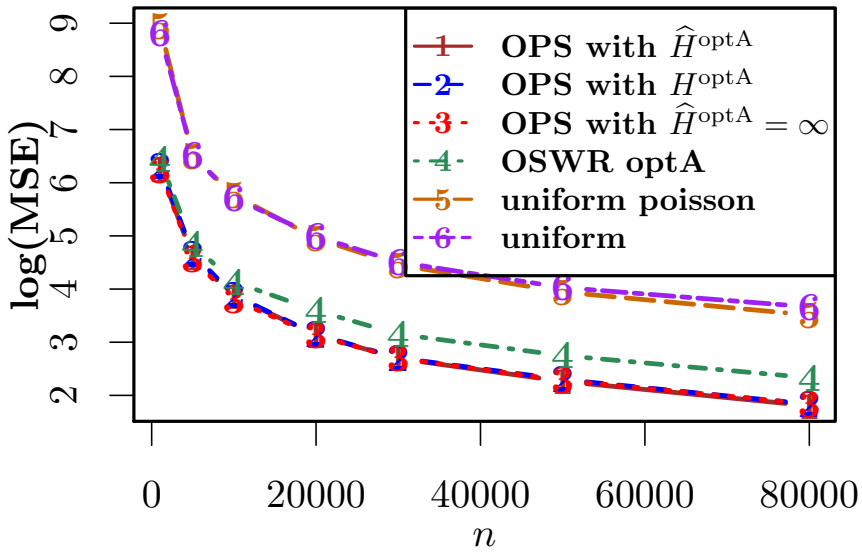


Figure 6 Empirical MSEs for cover type dataset among different n when $n_0 = 1000$ is fixed using different methods for 1000 replicates.

We assess the estimation efficiency of the approximately optimal Poisson subsampling algorithms by applying it to a forest cover type dataset [20]. This dataset is used to predict the forest type based on different geographical conditions. It contains 581012 observations and the response variable has 7 categories corresponding to 7 forest types whose percentages are 36.46% (Spruce/Fir), 48.76% (Lodgepole Pine), 6.15% (Ponderosa Pine), 0.427% (Cottonwood/Willow), 1.63% (Aspen), 2.99% (Douglas-fir) and 3.53% (Krummholz). We use the 10 quantitive variables as covariates, which measure geographical locations and lighting conditions. Figure 6 shows that the approximately optimal subsampling algorithms are more efficient than the algorithms based on uniform subsampling probabilities.

5.2.2 Character Font Image Data

We apply the approximately optimal subsampling algorithms to the character font image data [20]. This dataset contains pixel values and script information of different images for 153 fonts. We use observations from 5 fonts: Agency FB, Arial, Mongolian Baiti, Bank Gothic and

OCR-B as the target dataset and the number of total observations is 124817. The font type is the response variable and the corresponding percentages for each font are 0.80%, 21.02%, 1.32%, 1.79% and 75.06%. The dataset contains 410 quantitative variables and we use singular value decomposition to find 20 principle components with the largest singular values as the covariates. Figure 7 shows that the approximately optimal subsampling algorithms outperform the uniform subsampling algorithms.

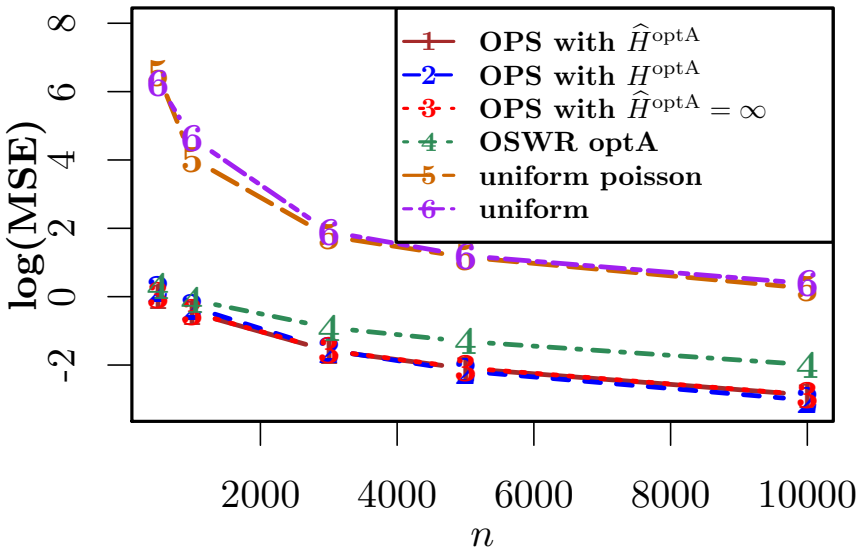


Figure 7 Empirical MSEs for character font image dataset among different n when $n_0 = 500$ is fixed using different methods for 1000 replicates.

6 Summary

In this paper, we have proposed an approximately optimal Poisson subsampling algorithm to reduce the computational burden in softmax regression. The Poisson subsampling procedure ensures that there are no duplicate data points in the subsample, and it is feasible to draw subsamples from massive datasets when the data volumes exceed the computer's memory limit. We have compared the proposed algorithm with the algorithm based on sampling with replacement on both simulated and real datasets, and demonstrated that the proposed algorithm has a better estimation efficiency, especially for high subsampling rate.

Acknowledgements We sincerely thank the Associate Editor and two anonymous reviewers for their comments, which greatly helped improve this paper.

References

[1] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

[2] P. Drineas, M. Magdon-Ismail, M.W. Mahoney, and D.P. Woodruff. Faster approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

[3] Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6):54:1–54:45, January 2017.

[4] Ping Ma, Michael W Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911, 2015.

[5] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152, Oct 2006.

[6] P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation. *Numerische Mathematik*, 117:219–249, 2011.

[7] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[8] HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2019.

[9] HaiYing Wang. Divide-and-conquer information-based optimalsubdata selection algorithm. *Journal of Statistical Theory and Practice*, 13(3):1–19, 2019.

[10] William Fithian and Trevor Hastie. Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics*, 42(5):1693, 2014.

[11] Lei Han, Kean Ming Tan, Ting Yang, Tong Zhang, et al. Local uncertainty sampling for large-scale multiclass logistic regression. *Annals of Statistics*, 48(3):1770–1788, 2020.

[12] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

[13] HaiYing Wang. More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132):1–59, 2019.

[14] Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31:749–772, 2021.

[15] Yaqiong Yao and HaiYing Wang. Optimal subsampling for softmax regression. *Statistical Papers*, 60(2):585–599, 2019.

[16] Guanyu Hu and HaiYing Wang. Most likely optimal subsampled Markov chain Monte Carlo. *Journal of Systems Science and Complexity*, 34(3):1121–1134, 2021.

[17] Haiying Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

[18] Jun Yu, HaiYing Wang, Mingyao Ai, and Huiming Zhang. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 0(0):1–12, 2020.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[20] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Supplementary Material

YAO Yaqiong · ZOU Jiahui · WANG HaiYing

A.1 Proofs Related to Optimal Poisson Subsampling Algorithms

To begin with, we introduce several notations. Denote $O_{P|\mathcal{D}_N}(1)$ and $o_{P|\mathcal{D}_N}(1)$ as boundedness and convergence to zero, respectively, in conditional probability given the full data $\mathcal{D}_N$. Specifically for a sequence of random vector $\mathbf{v}_{n,N}$, $\mathbf{v}_{n,N} = O_{P|\mathcal{D}_N}(1)$ means that for any $\varepsilon > 0$, there exists a finite $C_\varepsilon > 0$ such that

$$\mathbb{P}\left\{\sup_n \mathbb{P}\left(\|\mathbf{v}_{n,N}\| > C_\varepsilon|\mathcal{D}_N\right) \le \varepsilon\right\} \to 1 \quad \text{as} \quad n, N \to \infty;$$

$\mathbf{v}_{n,N} = o_{P|\mathcal{D}_N}(1)$ means that for any $\varepsilon, \delta > 0$,

$$\mathbb{P}\left\{\mathbb{P}(\|\mathbf{v}_{n,N}\| > \delta|\mathcal{D}_N) \le \varepsilon\right\} \to 1 \quad \text{as} \quad n, N \to \infty.$$

We use $\dot{f}(\boldsymbol{\beta})$ to denote the first derivative of function $f(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The asymptotic properties followed are obtained based on n and N tending to infinity except additional declarations.

Proof of Theorem 2.1

Proof (**Theorem 2.1**) Firstly, given the full data $\mathcal{D}_N$, under Assumptions 1 and 4, we have

$$\begin{aligned}\mathbb{E}\left\{\ell_p^*(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta})\big|\mathcal{D}_N\right\}^2 &= \frac{1}{N^2}\sum_{i=1}^N \frac{n\pi_i(1-n\pi_i)q_i^2(\boldsymbol{\beta})}{n^2\pi_i^2} \\ &\le \frac{1}{nN^2}\sum_{i=1}^N \frac{q_i^2(\boldsymbol{\beta})}{\pi_i}\end{aligned}$$

YAO Yaqiong
Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA.
Email: yaqiong.yao@uconn.edu
ZOU Jiahui
School of Statistics, Capital University of Economics and Business, Beijing 100070, China.
Email: zoujiahui@amss.ac.cn
WANG HaiYing (Corresponding author)
Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA.
Email: haiying.wang@uconn.edu

° YAO Yaqiong and ZOU Jiahui contributed equally to this work.

$$
\begin{aligned}
&\leq \frac{1}{nN^2}\sum_{i=1}^{N}\frac{2C_1^2\|\mathbf{x}_i\|^2+2C_2^2}{\pi_i}\\
&\leq \frac{2C_1^2}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^2}{\pi_i}+\frac{2C_2^2}{nN^2}\sum_{i=1}^{N}\frac{1}{\pi_i}\\
&= O_P(n^{-1}),
\end{aligned}
\tag{A.1}
$$

where $C_1=\lambda(K+1)$, $C_2=1+\log K$, $\lambda=\sup_{\boldsymbol{\beta}\in\Theta}\|\boldsymbol{\beta}\|$,

$$
\begin{aligned}
|q_i(\boldsymbol{\beta})| &= \left|\sum_{k=1}^{K}\delta_{i,k}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k-\log\left\{1+\sum_{l=1}^{K}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l}\right\}\right|\\
&\leq \sum_{k=1}^{K}\|\mathbf{x}_i\|\|\boldsymbol{\beta}_k\|+\log\left\{1+\sum_{l=1}^{K}e^{\|\mathbf{x}_i\|\|\boldsymbol{\beta}_l\|}\right\}\\
&\leq K\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|+\log\left(1+Ke^{\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|}\right)\\
&\leq K\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|+1+\log K+\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|\\
&\leq \lambda(K+1)\|\mathbf{x}_i\|+1+\log K\\
&= C_1\|\mathbf{x}_i\|+C_2
\end{aligned}
$$

and

$$
\frac{1}{N^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^2}{\pi_i}\leq\sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_i}}\sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i}}=O_P(1).
$$

From (A.1), by Markov's inequality, we have $\ell_P^*(\boldsymbol{\beta})-\ell_f(\boldsymbol{\beta})\to 0$ in conditional probability given $\mathcal{D}_N$. Note that the parameter space is compact, and $\widehat{\boldsymbol{\beta}}_{\text{sub}}^P$ and $\widehat{\boldsymbol{\beta}}_{\text{full}}$ are the unique global maximums of the continuous concave functions $\ell_P^*(\boldsymbol{\beta})$ and $\ell_f(\boldsymbol{\beta})$, respectively. Thus, from Theorem 5.9 and its remark of [1], conditionally on $\mathcal{D}_N$ in probability,

$$
\|\widehat{\boldsymbol{\beta}}_{\text{sub}}^P-\widehat{\boldsymbol{\beta}}_{\text{full}}\|=o_{P|\mathcal{D}_N}(1),
\tag{A.2}
$$

which ensures that $\widehat{\boldsymbol{\beta}}_{\text{sub}}^P$ is close to $\widehat{\boldsymbol{\beta}}_{\text{full}}$ as long as n is not small.

Secondly, using Taylor's theorem [c.f. Chapter 4 of 2],

$$
0=\dot{\ell}_{P,j}^*(\widehat{\boldsymbol{\beta}}_{\text{sub}}^P)=\dot{\ell}_{P,j}^*(\widehat{\boldsymbol{\beta}}_{\text{full}})+\frac{\partial\dot{\ell}_{P,j}^*(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}^T}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^P-\widehat{\boldsymbol{\beta}}_{\text{full}})+R_j^P,
\tag{A.3}
$$

where $\dot{\ell}_{P,j}^*(\boldsymbol{\beta})$ is the partial derivative of $\ell_P^*(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_j$, and

$$
R_j^P=(\widehat{\boldsymbol{\beta}}_{\text{sub}}^P-\widehat{\boldsymbol{\beta}}_{\text{full}})^T\int_0^1\int_0^1\frac{\partial^2\dot{\ell}_{P,j}^*\{\widehat{\boldsymbol{\beta}}_{\text{full}}+uv(\widehat{\boldsymbol{\beta}}_{\text{sub}}^P-\widehat{\boldsymbol{\beta}}_{\text{full}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}v\mathrm{d}u\mathrm{d}v\;(\widehat{\boldsymbol{\beta}}_{\text{sub}}^P-\widehat{\boldsymbol{\beta}}_{\text{full}}).
\tag{A.4}
$$

By direct calculation, the second derivative of $\dot{\ell}^*_{P,j}(\boldsymbol{\beta})$ satisfies the following condition

$$\sup_{\boldsymbol{\beta}\in\Theta}\left\|\frac{\partial^2\dot{\ell}^*_{P,j}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\|\leq\frac{2}{nN}\sum_{i=1}^{N}\frac{L\nu_i\|\mathbf{x}_i\|^3}{\pi_i}, \tag{A.5}$$

where L is a positive constant and $\|\cdot\|$ is the Frobenius norm. According to Markov's inequality and Assumption 3,

$$P\left(\frac{1}{nN}\sum_{i=1}^{N}\frac{\nu_i\|\mathbf{x}_i\|^3}{\pi_i}\geq\tau\middle|\mathcal{D}_N\right)\leq\frac{1}{nN\tau}\sum_{i=1}^{N}\mathbb{E}\left(\frac{\nu_i\|\mathbf{x}_i\|^3}{\pi_i}\middle|\mathcal{D}_N\right)=\frac{1}{N\tau}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3\to 0 \tag{A.6}$$

in probability as $\tau\to\infty$, which, combining with (A.5), deduces that

$$\sup_{u,v}\left\|\frac{\partial^2\dot{\ell}_{P,j}\{\widehat{\boldsymbol{\beta}}_{\text{full}}+uv(\widehat{\boldsymbol{\beta}}^P_{\text{sub}}-\widehat{\boldsymbol{\beta}}_{\text{full}})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right\|=O_{P|\mathcal{D}_N}(1), \tag{A.7}$$

and results in

$$R^P_j=O_{P|\mathcal{D}_N}(\|\widehat{\boldsymbol{\beta}}^P_{\text{sub}}-\widehat{\boldsymbol{\beta}}_{\text{full}}\|^2). \tag{A.8}$$

by (A.4). So, according to (A.3), we obtain

$$\widehat{\boldsymbol{\beta}}^P_{\text{sub}}-\widehat{\boldsymbol{\beta}}_{\text{full}}=-\ddot{\ell}^{*-1}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\left\{\dot{\ell}^*_P(\widehat{\boldsymbol{\beta}}_{\text{full}})+O_{P|\mathcal{D}_N}(\|\widehat{\boldsymbol{\beta}}^P_{\text{sub}}-\widehat{\boldsymbol{\beta}}_{\text{full}}\|^2)\right\}. \tag{A.9}$$

Thirdly, by direct calculation, we know that

$$\mathbb{E}\left\{\ddot{\ell}^*_P(\widehat{\boldsymbol{\beta}}_{\text{full}})|\mathcal{D}_N\right\}=\mathbf{M}_N. \tag{A.10}$$

For any component $[\ddot{\ell}^*_P]^{j_1j_2}$ of $\ddot{\ell}^*_P(\widehat{\boldsymbol{\beta}}_{\text{full}})$ and $1\leq j_1,j_2\leq dK$,

$$\begin{aligned}\mathbb{V}\left([\ddot{\ell}^*_P]^{j_1j_2}|\mathcal{D}_N\right)=&\frac{1}{(nN)^2}\sum_{i=1}^{N}\frac{n\pi_i(1-n\pi_i)\left([\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})]^{j_1j_2}\right)^2}{\pi_i^2}\\ \leq&\frac{1}{nN^2}\sum_{i=1}^{N}\frac{\left([\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})]^{j_1j_2}\right)^2}{\pi_i}\\ \leq&\frac{1}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_i}=O_P(n^{-1}),\end{aligned} \tag{A.11}$$

where the second last inequality holds by the fact that all elements of $\boldsymbol{\phi}_i$ are between 0 and 1, and the last equality is from Assumption 4. Combining with (A.10) and (A.11), we have

$$\ddot{\ell}^*_P(\widehat{\boldsymbol{\beta}}_{\text{full}})-\mathbf{M}_N=O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{A.12}$$

Afterwards, note that

$$\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\nu_i \mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes \mathbf{x}_i}{n\pi_i} \equiv \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\eta}_i^P. \tag{A.13}$$

Given $\mathcal{D}_N$, $\boldsymbol{\eta}_1^P, ..., \boldsymbol{\eta}_n^P$ are independent variables, we have

$$\mathbb{E}\{\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}})\} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes \mathbf{x}_i = 0 \tag{A.14}$$

and

$$\begin{aligned}
\mathbf{V}_{PG} &\equiv \mathbb{V}\left(\frac{\sqrt{n}}{N}\sum_{i=1}^{N}\boldsymbol{\eta}_i^P|\mathcal{D}_N\right) \qquad (A.15)\\
&= \frac{n}{N^2}\sum_{i=1}^{N}\frac{\mathbb{V}(\nu_i)\left\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes\mathbf{x}_i\right\}\left\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes\mathbf{x}_i\right\}^T}{(n\pi_i)^2}\\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{(1-n\pi_i)\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\pi_i}\\
&= O_P(1), \qquad (A.16)
\end{aligned}$$

where the last equality holds because each element of $\mathbf{V}_{PG}$ is bounded by $N^{-2}\sum_{i=1}^{N}\pi^{-1}\|\mathbf{x}_i\|^2 = O_P(1)$.

Meanwhile, for every $\varepsilon > 0$ and some $\rho > 0$,

$$\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\{\|\sqrt{n}N^{-1}\boldsymbol{\eta}_i^P\|^2 I(\sqrt{n}\|\boldsymbol{\eta}_i^P\| > N\varepsilon)|\mathcal{D}_N\}\\
&\le \frac{n^{1+\rho/2}}{N^{2+\rho}\varepsilon^{\rho}}\sum_{i=1}^{N}\mathbb{E}\{\|\boldsymbol{\eta}_i^P\|^{2+\rho} I(\sqrt{n}\|\boldsymbol{\eta}_i^P\| > N\varepsilon)|\mathcal{D}_N\}\\
&\le \frac{n^{1+\rho/2}}{N^{2+\rho}\varepsilon^{\rho}}\sum_{i=1}^{N}\mathbb{E}(\|\boldsymbol{\eta}_i^P\|^{2+\rho}|\mathcal{D}_N)\\
&= \frac{n^{1+\rho/2}}{N^{2+\rho}\varepsilon^{\rho}}\sum_{i=1}^{N}\frac{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|^{2+\rho}\|\mathbf{x}_i\|^{2+\rho}}{(n\pi_i)^{1+\rho}}\\
&\le \frac{1}{n^{\rho/2}}\frac{1}{N^{2+\rho}}\frac{1}{\varepsilon^{\rho}}\sum_{i=1}^{N}\frac{K^{2+\rho}\|\mathbf{x}_i\|^{2+\rho}}{\pi_i^{1+\rho}} = o_P(1),
\end{aligned}$$

where the last equality is from Assumption 4. From (A.13) and (A.16), by the Lindeberg-Feller

central limit theorem [Proposition 2.27 of 1],

$$\sqrt{n}\mathbf{V}_{PG}^{-1/2}\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \mathbf{V}_{PG}^{-1/2}\frac{\sqrt{n}}{N}\sum_{i=1}^{N}\boldsymbol{\eta}_i^P \to \mathbb{N}(0,\mathbf{I}) \tag{A.17}$$

in distribution conditionally on $\mathcal{D}_N$.

Finally, from (A.2), (A.9) and (A.12), we have

$$\widehat{\boldsymbol{\beta}}_{\text{sub}}^P - \widehat{\boldsymbol{\beta}}_{\text{full}} = -\ddot{\ell}_P^{*-1}(\widehat{\boldsymbol{\beta}}_{\text{full}})\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) + o_{P|\mathcal{D}_N}(1), \tag{A.18}$$

$$\ddot{\ell}_P^{*-1}(\widehat{\boldsymbol{\beta}}_{\text{full}}) - \mathbf{M}_N^{-1} = -\mathbf{M}_N^{-1}\{\ddot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) - \mathbf{M}_N\}\ddot{\ell}_P^{*-1}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = O_{P|\mathcal{D}_N}(n^{-1/2}) \tag{A.19}$$

and

$$\mathbf{V}_G^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{PG}^{1/2}(\mathbf{V}_G^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{PG}^{1/2})^T = \mathbf{V}_G^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{PG}^{1/2}\mathbf{V}_{PG}^{1/2}\mathbf{M}_N^{-1}\mathbf{V}_G^{-1/2} = \mathbf{I}. \tag{A.20}$$

Then gathering (A.17), (A.18), (A.19) and (A.20), by Slutsky's Theorem [Theorem 6 of 2], we have

$$\begin{aligned}
&\sqrt{n}\mathbf{V}_G^{-1/2}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^P - \widehat{\boldsymbol{\beta}}_{\text{full}}) \\
&= -\sqrt{n}\mathbf{V}_G^{-1/2}\ddot{\ell}_P^{*-1}(\widehat{\boldsymbol{\beta}}_{\text{full}})\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) + o_{P|\mathcal{D}_N}(1) \\
&= -\sqrt{n}\mathbf{V}_G^{-1/2}\mathbf{M}_N^{-1}\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) - \sqrt{n}\mathbf{V}_G^{-1/2}\{\ddot{\ell}_P^{*-1}(\widehat{\boldsymbol{\beta}}_{\text{full}}) - \mathbf{M}_N^{-1}\}\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) + o_{P|\mathcal{D}_N}(1) \\
&= -\mathbf{V}_G^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{PG}^{1/2}\sqrt{n}\mathbf{V}_{PG}^{-1/2}\dot{\ell}_P^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) + o_{P|\mathcal{D}_N}(1) \\
&\to \mathbb{N}(0,\mathbf{I})
\end{aligned}$$

in distribution conditionally on $\mathcal{D}_N$, where $\mathbf{V}_G = \mathbf{M}_N^{-1}\mathbf{V}_{PG}\mathbf{M}_N^{-1}$.

Proof of Theorem 3.1

Proof (**Theorem 3.1**) We only prove this theorem under A-optimality criterion because the proof under L-optimality criterion is similar. The optimal subsampling probabilities under A-optimality criterion minimize $\text{tr}(\mathbf{V}_G)$, which is

$$\begin{aligned}
\text{tr}(\mathbf{V}_G) &= \text{tr}(\mathbf{M}_N^{-1}\mathbf{V}_{PG}\mathbf{M}_N^{-1}) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-n\pi_i}{\pi_i}\text{tr}\left\{M_N^{-1}\psi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\text{T}})M_N^{-1}\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-n\pi_i}{\pi_i}\|M_N^{-1}\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes\mathbf{x}_i\}\|^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{(t_i^{\text{optA}})^2}{\pi_i} - \frac{n}{N^2}\sum_{i=1}^{N}(t_i^{\text{optA}})^2.
\end{aligned}$$

For simplicity, let $t_i = t_{(i)}^{\text{optA}}(i = 1, 2, ..., N)$. This optimization problem can be reduced to

find π_i^{optA} minimizing

$$T = \sum_{i=1}^{N} \frac{t_i^2}{\pi_i} \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1, \quad 0 \le \pi_i \le \frac{1}{n}, \quad i = 1, 2, ..., N. \tag{A.21}$$

Utilizing slack variables $\omega_1^2, \omega_2^2, ..., \omega_N^2$ and by Lagrangian multiplier method, we can construct

$$H(\pi_1, ..., \pi_N, \lambda, \mu_1, ..., \mu_N, \omega_1, ..., \omega_N) = \sum_{i=1}^{N} \frac{t_i^2}{\pi_i} + \lambda \left(\sum_{i=1}^{N} \pi_i - 1 \right) + \sum_{i=1}^{N} \mu_i \left(\pi_i + \omega_i^2 - \frac{1}{n} \right). \tag{A.22}$$

The KKT conditions [3]

$$\begin{cases} \dfrac{\partial H}{\partial \pi_i} = -\dfrac{t_i^2}{\pi_i^2} + \lambda + \mu_i = 0, & i = 1, 2, ..., N; & \text{(A.23)} \\ \dfrac{\partial H}{\partial \lambda} = \displaystyle\sum_{i=1}^{N} \pi_i - 1 = 0; & & \text{(A.24)} \\ \dfrac{\partial H}{\partial \mu_i} = \pi_i + \omega_i^2 = \dfrac{1}{n}, & i = 1, 2, ..., N; & \text{(A.25)} \\ \dfrac{\partial H}{\partial \omega_i} = 2\mu_i \omega_i = 0, & i = 1, 2, ..., N; & \text{(A.26)} \\ \mu_i \ge 0, & i = 1, 2, ..., N. & \text{(A.27)} \end{cases}$$

are satisfied. According to (A.23), we have

$$\pi_i = \frac{t_i}{\sqrt{\lambda + \mu_i}}, i = 1, 2, ..., N. \tag{A.28}$$

Combined with (A.25),

$$\frac{t_i}{\sqrt{\lambda + \mu_i}} + \omega_i^2 = \frac{1}{n}, i = 1, 2, ..., N; \tag{A.29}$$

According to (A.26), at least one of μ_i and ω_i should be 0. Then we have the following equations

$$t_i \le \frac{\sqrt{\lambda}}{n}, \quad \mu_i = 0, \quad \pi_i = \frac{t_i}{\sqrt{\lambda}}; \tag{A.30}$$

$$t_i > \frac{\sqrt{\lambda}}{n}, \quad \omega_i = 0, \quad \pi_i = \frac{t_i}{\sqrt{\lambda + \mu_i}} = \frac{1}{n}. \tag{A.31}$$

Here t_i is organized in an increasing order, and let $t_{N-g} \le \frac{\sqrt{\lambda}}{n}$ and $t_{N-g+1} > \frac{\sqrt{\lambda}}{n}$ $(g \ge 0)$. From

(A.24), we have

$$\sum_{i=1}^{N-g} \frac{t_i}{\sqrt{\lambda}} + \sum_{i=N-g+1}^{N} \frac{t_i}{\sqrt{\lambda+\mu_i}} = 1,$$
$$\sum_{i=1}^{N-g} \frac{t_i}{\sqrt{\lambda}} + \sum_{i=N-g+1}^{N} \frac{1}{n} = 1.$$

Thus,

$$\sqrt{\lambda} = \frac{n}{n-g} \sum_{i=1}^{N-g} t_i. \tag{A.32}$$

Let $H = \frac{\sqrt{\lambda}}{n} = \frac{\sum_{i=1}^{N-g} t_i}{n-g}$, then we have

$$\begin{aligned}\sum_{i=1}^{N}(t_i \wedge H) &= \sum_{i=1}^{N-g} t_i + \sum_{i=N-g+1}^{N} H \\ &= \sum_{i=1}^{N-g} t_i + \frac{g}{n-g} \sum_{i=1}^{N-g} t_i \\ &= \sqrt{\lambda} = nH. \end{aligned} \tag{A.33}$$

Combining with (A.31) and (A.30), we know

$$\pi_i = \frac{t_i}{\sum_{i=1}^{N}(t_i \wedge H)}, t_i < H;$$
$$\pi_i = \frac{H}{\sum_{i=1}^{N}(t_i \wedge H)}, t_i \geq H.$$

This can be further simplified as

$$\pi_i = \frac{t_i \wedge H}{\sum_{i=1}^{N}(t_i \wedge H)}, i = 1, 2, ..., N. \tag{A.34}$$

The range of g is shown as follows. Combined with (A.33) and (A.34), we know

$$\begin{cases} \pi_{N-g} = \dfrac{t_{N-g}(n-g)}{n\sum_{i=1}^{N-g} t_i} < \dfrac{1}{n}, \\ \pi_{N-g+1} = \dfrac{1}{n} \geq \dfrac{t_{N-g+1}(n-g+1)}{n\sum_{i=1}^{N-g+1} t_i}; \end{cases}$$

$$\Leftrightarrow \begin{cases} \dfrac{t_{N-g}}{\sum_{i=1}^{n-g} t_i} < \dfrac{1}{n-g}, \\ \dfrac{t_{N-g+1}}{\sum_{i=1}^{N-g+1} t_i} \geq \dfrac{1}{n-g+1}; \end{cases}$$

$$\Leftrightarrow \begin{cases} \dfrac{\sum_{i=1}^{N-g} t_i}{t_{N-g}} > n-g, \\ \dfrac{\sum_{i=1}^{N-g+1} t_i}{t_{N-g+1}} \leq n-g+1. \end{cases}$$

Proof of Theorem 4.3

For clear presentation, we use $\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})$ to represent π_i^{optL} and use $\pi_i(\widehat{\boldsymbol{\beta}}_P^0)$ to represent the quantity with same expression as π_i^{optL} except that $\widehat{\boldsymbol{\beta}}_{\text{full}}$ is replaced by $\widehat{\boldsymbol{\beta}}_P^0$. The sample sizes of the two stages are set to be $n_0 = o(n^{1/2})$ to ensure that the contribution of the first stage subsample is dominated by that of the second stage subsample. Thus, we only need to consider the second stage subsample in the objective function. Denote

$$\begin{aligned} \ell_P^{*ada}(\boldsymbol{\beta}) &= \frac{1}{N}\sum_{i=1}^{n^*} \frac{1}{n\pi_i^*(\widehat{\boldsymbol{\beta}}_P^0)} \left[\sum_{k=1}^{K} \delta_{i,k}^* \boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{x}_i^* - \log\left\{1 + \sum_{l=1}^{K} e^{\boldsymbol{\beta}_l^{\mathrm{T}} \mathbf{x}_i^*}\right\}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{\nu_i}{n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \left[\sum_{k=1}^{K} \delta_{i,k} \boldsymbol{\beta}_k^{T} \mathbf{x}_i - \log\left\{1 + \sum_{l=1}^{K} e^{\boldsymbol{\beta}_l^{T} \mathbf{x}_i}\right\}\right], \end{aligned}$$

where $\nu_i \sim Bern\{n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)\}$ is the indicator variable in the second stage. For sake of simplicity, we use * to denote quantities of the second stage sample in this proof.

Before proofing, we need several lemmas.

Lemma A.1.1 *Under Assumption 5, as* $N \to \infty$

$$\|\mathbf{x}\|_{(N)} = O_P(\ln N),$$

where $\|\mathbf{x}\|_{(N)} = \max\{\|\mathbf{x}_i\|, i = 1, 2, ..., N\}$.

Proof (**Lemma A.1.1**) Note that for some $M > 1$ and a constant $c > 0$,

$$\lim_{N\to\infty} N \ln\left\{1 - \frac{\exp(c^2/2)}{N^M}\right\} = \lim_{N\to\infty} -\frac{\exp(c^2/2)}{N^{M-1}} = 0 \tag{A.35}$$

Thus, combing (A.35) with Markov's inequality and Assumption 5, we know that for any $\varepsilon > 0$, there exists a sufficient large N such that

$$\begin{aligned} \mathbb{P}\{\|\mathbf{x}\|_{(N)} \leq M \ln N\} &= \{1 - \mathbb{P}(\|\mathbf{x}_i\| > M \ln N)\}^N \\ &\geq \left\{1 - \frac{\mathbb{E}(e^{\|\mathbf{x}_i\|})}{e^{M \ln N}}\right\}^N \end{aligned}$$

$$
\begin{aligned}
&\geq \left\{1 - \frac{\exp(c^2/2)}{N^M}\right\}^N \\
&= \exp\left[N \ln\left\{1 - \frac{\exp(c^2/2)}{N^M}\right\}\right] \\
&= 1 - \varepsilon.
\end{aligned}
$$

Therefore,

$$
\|\mathbf{x}\|_{(N)} = O_P(\ln N).
$$

Lemma A.1.2 *Under Assumptions 1 and 5, if $k_2 \geq 1$ and $k_1 - k_2 \geq -1$, then*

$$
\begin{aligned}
\frac{1}{N^{k_2+1}} \sum_{i=1}^N \frac{\|\mathbf{x}_i\|^{k_1}}{\pi_i^{k_2}(\widehat{\boldsymbol{\beta}}_P^0)} \leq K^{k_2/2} &\left(\frac{Kn}{n-g}\frac{1}{N}\sum_{j=1}^N \|\mathbf{x}_j\|\right)^{k_2} \left(\frac{1}{N}\sum_{j=1}^N \|\mathbf{x}_j\|^{k_1-k_2}\left(1 + Ke^{\lambda\|\mathbf{x}_j\|}\right)^{k_2}\right) \\
&+ \frac{n^{k_2}}{N^{k_2}}\left(\frac{1}{N}\sum_{j=1}^N \|\mathbf{x}_i\|^{k_1}\right) = O_P(1),
\end{aligned}
$$

where $\lambda = \sup_{\boldsymbol{\beta}\in\Theta} \|\boldsymbol{\beta}\|$.

Proof (**Lemma A.1.2**) For simplicity, here $\{t_i\}_{i=1}^N$ means $\{t_i^{\mathrm{optL}}(\widehat{\boldsymbol{\beta}}_P^0)\}_{i=1}^N$, whose expression is the same as (7) except replacing $\widehat{\boldsymbol{\beta}}_{\mathrm{full}}$ with $\widehat{\boldsymbol{\beta}}_P^0$. Denote the order statistic of $\{t_i\}_{i=1}^N$ as $\{t_{(i)}\}_{i=1}^N$ and reorder $\{\mathbf{x}_i\}_{i=1}^N$, $\{y_i\}_{i=1}^N$ and $\{\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_P^0)\}_{i=1}^N$ to be the same sequence as $\{t_{(i)}\}_{i=1}^N$ and define them to be $\{\mathbf{x}_i'\}_{i=1}^N$, $\{y_i'\}_{i=1}^N$ and $\{\mathbf{s}_i'(\widehat{\boldsymbol{\beta}}_P^0)\}_{i=1}^N$, respectively.

Firstly, it is seen that

$$
\begin{aligned}
\frac{1}{N^{k_2+1}} \sum_{i=1}^N \frac{\|\mathbf{x}_i\|^{k_1}}{\pi_i^{k_2}(\widehat{\boldsymbol{\beta}}_P^0)} &= \frac{1}{N^{k_2+1}}\left\{\sum_{i=1}^N (t_i \wedge H)\right\}^{k_2}\left\{\sum_{i=1}^N \frac{\|\mathbf{x}_i\|^{k_1}}{(t_i \wedge H)^{k_2}}\right\} \\
&= \frac{1}{N^{k_2+1}}\left\{\sum_{i=1}^N (t_i \wedge H)\right\}^{k_2}\left[\sum_{i=1}^N \frac{\|\mathbf{x}_i'\|^{k_1}}{\{t_{(i)} \wedge H\}^{k_2}}\right] \\
&= \frac{1}{N^{k_2+1}}\left\{\sum_{i=1}^N (t_i \wedge H)\right\}^{k_2}\left[\sum_{i=1}^{N-g} \frac{\|\mathbf{x}_i'\|^{k_1}}{\{t_{(i)} \wedge H\}^{k_2}} + \sum_{i=N-g+1}^N \frac{\|\mathbf{x}_i'\|^{k_1}}{\{t_{(i)} \wedge H\}^{k_2}}\right] \\
&= \frac{1}{N^{k_2+1}}\left\{\sum_{i=1}^N (t_i \wedge H)\right\}^{k_2}\left\{\sum_{i=1}^{N-g} \frac{\|\mathbf{x}_i'\|^{k_1}}{t_{(i)}^{k_2}} + \sum_{i=N-g+1}^N \frac{\|\mathbf{x}_i'\|^{k_1}}{H^{k_2}}\right\} \\
&= \frac{1}{N^{k_2+1}}\left\{\sum_{i=1}^N (t_i \wedge H)\right\}^{k_2}\left\{\sum_{i=1}^{N-g} \frac{\|\mathbf{x}_i'\|^{k_1-k_2}}{\|\mathbf{s}_i'(\widehat{\boldsymbol{\beta}}_P^0)\|^{k_2}} + \sum_{i=N-g+1}^N \frac{\|\mathbf{x}_i'\|^{k_1}}{H^{k_2}}\right\} \\
&\equiv \Delta_1 + \Delta_2. \qquad \text{(A.36)}
\end{aligned}
$$

Secondly, we have

$$\begin{aligned}
\|\mathbf{s}_i'(\widehat{\boldsymbol{\beta}}_P^0)\| &= \left[\sum_{k=1}^{K}\{\delta_{i,k} - p_k(\mathbf{x}_i', \boldsymbol{\beta})\}^2\right]^{1/2} \\
&\geq K^{-\frac{1}{2}} \sum_{k=1}^{K} |\delta_{i,k} - p_k(\mathbf{x}_i', \boldsymbol{\beta})| \\
&= K^{-\frac{1}{2}} \left\{ \frac{1 + \sum_{l=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_l} I(l \neq j)}{1 + \sum_{k=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_k}} + \frac{\sum_{l=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_l} I(l \neq j)}{1 + \sum_{k=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_k}} \right\} \\
&= K^{-\frac{1}{2}} \left\{ \frac{1 + 2\sum_{l=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_l}(1 - \delta_{i,l})}{1 + \sum_{k=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_k}} \right\} \\
&\geq K^{-\frac{1}{2}} \frac{1}{1 + \sum_{k=1}^{K} e^{\mathbf{x}_i'^T \boldsymbol{\beta}_k}} \\
&\geq K^{-\frac{1}{2}} \left(1 + K e^{\lambda \|\mathbf{x}_i'\|}\right)^{-1}, \qquad \text{(A.37)}
\end{aligned}$$

where the first inequality is due to Cauchy-Schwartz inequality. Note that from Assumption 5 and Law of Large Numbers, for a given $k > 0$, we know

$$\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i\|^k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\|\mathbf{x}_i\|^k + o_P(1) = O_P(1), \qquad \text{(A.38)}$$

and for a given $a > 0$,

$$\frac{1}{N}\sum_{i=1}^{N} e^{a\|\mathbf{x}_i\|} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}e^{a\|\mathbf{x}_i\|} + o_P(1) = O_P(1). \qquad \text{(A.39)}$$

Then, by (A.38), (A.39) and Assumption 5, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(1 + K e^{\lambda\|\mathbf{x}_i\|}\right)^{k_2} \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(2K e^{\lambda\|\mathbf{x}_i\|}\right)^{k_2} \\
&\leq (2K)^{k_2} \left\{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i\|^{2(k_1 - k_2)}\right\}^{1/2} \left(\frac{1}{N}\sum_{i=1}^{N} e^{2k_2\lambda\|\mathbf{x}_i\|}\right)^{1/2} = O_P(1), \qquad \text{(A.40)}
\end{aligned}$$

where the last inequality is according to Cauchy-Schwartz inequality.

Finally, with assumption 5, (A.37), (A.38) and (A.40), we can achieve

$$\Delta_1 = \frac{1}{N^{k_2+1}} \left\{\sum_{j=1}^{N} (t_j \wedge H)\right\}^{k_2} \sum_{i=1}^{N-g} \frac{\|\mathbf{x}_i'\|^{k_1 - k_2}}{\|\mathbf{s}_i'(\widehat{\boldsymbol{\beta}}_P^0)\|^{k_2}}$$

$$
\begin{aligned}
&\leq K^{k_2/2}\left(\frac{n}{n-g}\frac{1}{N}\sum_{i=1}^{N-g}t_{(i)}\right)^{k_2}\left\{\frac{1}{N}\sum_{i=1}^{N-g}\|\mathbf{x}_i'\|^{k_1-k_2}\left(1+Ke^{\lambda\|\mathbf{x}_i'\|}\right)^{k_2}\right\}\\
&\leq K^{k_2/2}\left(\frac{n}{n-g}\frac{1}{N}\sum_{i=1}^{N-g}\|\mathbf{s}_i'(\widehat{\boldsymbol{\beta}}_P^0)\|\|\mathbf{x}_i'\|\right)^{k_2}\left\{\frac{1}{N}\sum_{i=1}^{N-g}\|\mathbf{x}_i'\|^{k_1-k_2}\left(1+Ke^{\lambda\|\mathbf{x}_i'\|}\right)^{k_2}\right\}\\
&\leq K^{k_2/2}\left(\frac{\sqrt{K}n}{n-g}\frac{1}{N}\sum_{i=1}^{N-g}\|\mathbf{x}_i'\|\right)^{k_2}\left\{\frac{1}{N}\sum_{i=1}^{N-g}\|\mathbf{x}_i'\|^{k_1-k_2}\left(1+Ke^{\lambda\|\mathbf{x}_i'\|}\right)^{k_2}\right\}\\
&\leq K^{k_2/2}\left(\frac{\sqrt{K}n}{n-g}\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|\right)^{k_2}\left\{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^{k_1-k_2}\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)^{k_2}\right\}\\
&= O_P(1). \qquad \text{(A.41)}
\end{aligned}
$$

With the help of (A.33) and (A.38), we have

$$
\begin{aligned}
\Delta_2 &= \frac{1}{N^{k_2+1}}\left\{\frac{\sum_{j=1}^{N}(t_j\wedge H)}{H}\right\}^{k_2}\sum_{i=N-g+1}^{N}\|\mathbf{x}_i'\|^{k_1}\\
&= \frac{n^{k_2}}{N^{k_2+1}}\sum_{i=N-g+1}^{N}\|\mathbf{x}_i'\|^{k_1}\\
&\leq \frac{n^{k_2}}{N^{k_2}}\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^{k_1}\right)\\
&= O_P\left(\frac{n^{k_2}}{N^{k_2}}\right). \qquad \text{(A.42)}
\end{aligned}
$$

This proof is completed combining with (A.36), (A.41) and (A.42).

Lemma A.1.3 *If Assumptions 1, 2 and 5 hold, then*

$$
\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})-\mathbf{M}_N = O_{P|\mathcal{D}_N}(n^{-1/2}) \qquad \text{(A.43)}
$$

and

$$
\dot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = O_{P|\mathcal{D}_N}(n^{-1/2}), \qquad \text{(A.44)}
$$

where

$$
\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \frac{\partial^2\ell_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{1}{nN}\sum_{i=1}^{N}\frac{\nu_i\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i{}^{\mathrm{T}})}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}.
$$

Proof (**Lemma A.1.3**) Calculate directly,

$$
\mathbb{E}\left\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})|\mathcal{D}_N\right\} = \mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}\left[\mathbb{E}\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}_P^0\}\right]
$$

$$= \mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}(\mathbf{M}_N|\mathcal{D}_N) = \mathbf{M}_N, \tag{A.45}$$

where $\mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}$ means the expectation is taken with respect to the distribution of $\widehat{\boldsymbol{\beta}}_P^0$ conditionally on $\mathcal{D}_N$.

For any element $\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\}^{j_1 j_2}$ of $\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})$ and $1 \le j_1, j_2 \le dK$,

$$\begin{aligned}\mathbb{V}\left[\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\}^{j_1 j_2}|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0\right] &= \frac{1}{nN^2}\sum_{i=1}^{N}\frac{\left\{1-n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)\right\}\left[\{\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\}^{j_1 j_2}\right]^2}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \\ &\le \frac{1}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)},\end{aligned} \tag{A.46}$$

where the last inequality holds by the fact that all elements of $\boldsymbol{\phi}_i$ are between 0 and 1. Further, from Lemma A.1.2 and (A.46),

$$\begin{aligned}&\mathbb{V}\left\{[\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})]^{j_1 j_2}|\mathcal{D}_N\right\} \\ &= \mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}\left(\mathbb{V}\left[\left\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{j_1 j_2}|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0\right]\right) \\ &\quad + \mathbb{V}_{\widehat{\boldsymbol{\beta}}_P^0}\left(\mathbb{E}\left[\left\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{j_1 j_2}|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0\right]\right) \\ &= \mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}\left(\mathbb{V}\left[\left\{\ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{j_1 j_2}|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0\right]\right) \\ &\le \mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}\left\{\frac{1}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}\right\} \\ &\le \mathbb{E}_{\widehat{\boldsymbol{\beta}}_P^0}\left\{\frac{K^{1/2}}{n}\left(\frac{Kn}{n-g}\frac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_j\|\right)\left(\frac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_j\|^3\left(1+Ke^{\lambda\|\mathbf{x}_j\|}\right)\right)\right. \\ &\quad \left. + \frac{1}{nN}\left(\frac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_i\|^4\right)\right\} \\ &= O_P(n^{-1}).\end{aligned} \tag{A.47}$$

Using Markov's inequality, (A.43) follows from (A.45) and (A.47).

Similarly, we can achieve that

$$\mathbb{E}\left\{\dot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\middle|\mathcal{D}_N\right\} = 0, \tag{A.48}$$

$$\mathbb{V}\left\{\dot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})\middle|\mathcal{D}_N\right\} = O_P(n^{-1}). \tag{A.49}$$

Combing with (A.48), (A.49) and Markov's inequality, (A.44) is obtained.

Lemma A.1.4 *If Assumptions 1, 2 and 5 hold, then*

$$\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}} = O_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0}(n^{-1/2}). \tag{A.50}$$

Proof (**Lemma A.1.4**) Firstly, for any $\boldsymbol{\beta} \in \Theta$, we have

$$\begin{aligned}
\mathbb{E}\left\{\ell_P^{*ada}(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta}) \middle| \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0\right\}^2 &= \frac{1}{n}\left[\frac{1}{N^2}\sum_{i=1}^{N}\frac{\{1 - n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)\}q_i^2(\boldsymbol{\beta})}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}\right] \\
&\leq \frac{1}{n}\left[\frac{1}{N^2}\sum_{i=1}^{N}\frac{2C_1^2\|\mathbf{x}_i\|^2 + 2C_2^2}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}\right] \\
&= O_P(n^{-1}),
\end{aligned} \tag{A.51}$$

where $C_1 = \lambda(K+1)$, $C_2 = 1 + \log K$ and the last inequality is due to Lemma A.1.2.

Now, according to (A.51), we have $\ell_P^{*ada}(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta}) \to 0$ in conditional probability conditionally on $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}_P^0$. Note that the parameter space is compact, and $\widehat{\boldsymbol{\beta}}_P^{ada}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}}$ are the global maximus of the continuous concave functions $\ell_P^{*ada}(\boldsymbol{\beta})$ and $\ell_f(\boldsymbol{\beta})$, respectively. Thus,

$$\|\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}}\| = o_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0}(1). \tag{A.52}$$

Secondly, using Taylor's theorem [c.f. Chapter 4 of 2],

$$0 = \dot{\ell}_{P,j}^{*ada}(\widehat{\boldsymbol{\beta}}_P^{ada}) = \dot{\ell}_{P,j}^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) + \frac{\partial \dot{\ell}_{P,j}^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial \boldsymbol{\beta}^T}(\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}}) + R_j^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}, \tag{A.53}$$

where $\dot{\ell}_{P,j}^{*ada}(\boldsymbol{\beta})$ is the partial derivative of $\ell_P^{*ada}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_j$, and

$$R_j^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0} = (\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}})^T \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_{P,j}^{*ada}\{\widehat{\boldsymbol{\beta}}_{\text{full}} + uv(\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v \mathrm{d}u \mathrm{d}v \ (\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}}). \tag{A.54}$$

The second derivative of $\dot{\ell}_{P,j}^{*ada}(\boldsymbol{\beta})$ satisfies the following condtion

$$\sup_{\boldsymbol{\beta} \in \Theta} \left\| \frac{\partial^2 \dot{\ell}_{P,j}^{*ada}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\| \leq \frac{2}{n}\sum_{i=1}^{N} \frac{L' \nu_i \|\mathbf{x}_i\|^3}{N\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} = O_{P|\mathcal{D}_N}(n^{-1}), \tag{A.55}$$

where L' is a positive constant and the last equality is due to

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{N}\frac{\nu_i\|\mathbf{x}_i\|^3}{N\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \geq \tau \middle| \mathcal{D}_N\right\} \leq \frac{1}{nN\tau}\sum_{i=1}^{N}\mathbb{E}\left\{\frac{\nu_i\|\mathbf{x}_i\|^3}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \middle| \mathcal{D}_N\right\} = \frac{1}{N\tau}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3 \to 0 \tag{A.56}$$

in probability as $\tau \to \infty$. So, from (A.55) we have

$$\sup_{u,v} \left\| \frac{\partial^2 \dot{\ell}_{P,j}^{*ada}\{\widehat{\boldsymbol{\beta}}_{\text{full}} + uv(\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\| = O_{P|\mathcal{D}_N}(1), \tag{A.57}$$

which, combing with (A.54), deduces

$$R_j^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0} = O_{P|\mathcal{D}_N}\left(\|\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\|^2 \right). \tag{A.58}$$

Finally, from (A.53) and (A.58),

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}} &= -\left\{ \ddot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) \right\}^{-1} \left\{ \dot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N}(\|\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\|^2) \right\} \\ &= O_{P|\mathcal{D}_N}(n^{-1/2}) + o_{P|\mathcal{D}_N}(\|\widehat{\boldsymbol{\beta}}_P^{ada} - \widehat{\boldsymbol{\beta}}_{\text{full}}\|) \\ &= o_{P|\mathcal{D}_N}(1), \end{aligned} \tag{A.59}$$

where the second equality is from (A.43) in Lemma A.1.3 and the last equality is due to (A.52).

Proof (**Theorem 4.3**) Firstly, denote

$$\dot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \frac{1}{N} \sum_{i=1}^N \frac{\nu_i \mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i}{n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \equiv \frac{1}{N} \sum_{i=1}^N \boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0}. \tag{A.60}$$

Given $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}_P^0$, $\boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0} (i = 1, 2, ..., n)$ are independent. We also have that

$$\begin{aligned} \mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}_P^0} &\equiv \mathbb{V}\left(\frac{\sqrt{n}}{N} \sum_{i=1}^N \boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0} \Big| \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0 \right) \\ &= \frac{1}{nN^2} \sum_{i=1}^N \frac{\mathbb{V}(\nu_i) \left\{ \mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i \right\} \left\{ \mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i \right\}^T}{\pi_i^2(\widehat{\boldsymbol{\beta}}_P^0)} \\ &= \frac{1}{N^2} \sum_{i=1}^N \frac{\left\{ 1 - n\pi_i(\widehat{\boldsymbol{\beta}}_P^0) \right\} \left\{ \boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i \mathbf{x}_i^T) \right\}}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \\ &= O_P(1), \end{aligned} \tag{A.61}$$

where the last equality holds because each element of $\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}_P^0}$ is bounded by $N^{-2} \sum_{i=1}^N \pi_i(\widehat{\boldsymbol{\beta}}_P^0)^{-1} \|\mathbf{x}_i\|^2$ and $N^{-2} \sum_{i=1}^N \pi_i(\widehat{\boldsymbol{\beta}}_P^0)^{-1} \|\mathbf{x}_i\|^2 = O_P(1)$ from Lemma A.1.2.

Meanwhile, for every $\varepsilon > 0$ and some $\rho > 0$,

$$\sum_{i=1}^N \mathbb{E}\{ \|\sqrt{n} N^{-1} \boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0}\|^2 I(\|\boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0}\| > n^{-1/2} N\varepsilon) | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0 \}$$

$$
\begin{aligned}
&\leq \frac{n^{1+\rho/2}}{N^{1+\rho/2}\varepsilon^{\rho}} \sum_{i=1}^{N} \mathbb{E}\{\|\boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0}\|^{2+\rho} I(\|\boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0}\| > n^{1/2}\varepsilon) | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0\} \\
&\leq \frac{n^{1+\rho/2}}{N^{1+\rho/2}\varepsilon^{\rho}} \sum_{i=1}^{N} \mathbb{E}(\|\boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0}\|^{2+\rho} | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0) \\
&= \frac{n^{1+\rho/2}}{N^{1+\rho/2}\varepsilon^{\rho}} \sum_{i=1}^{N} \frac{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|^{2+\rho} \|\mathbf{x}_i\|^{2+\rho}}{\{n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)\}^{1+\rho}} \\
&\leq \frac{1}{n^{\rho/2}} \frac{1}{N^{2+\rho}} \frac{1}{\varepsilon^{\rho}} \sum_{i=1}^{N} \frac{K^{2+\rho}\|\mathbf{x}_i\|^{2+\rho}}{\pi_i^{1+\rho}(\widehat{\boldsymbol{\beta}}_P^0)} = o_P(1),
\end{aligned}
$$

where the last equality is from Lemma A.1.2. From (A.60) and (A.61), by the Lindeberg-Feller central limit theorem [Proposition 2.27 of 1], conditionally on $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}_P^0$,

$$
\sqrt{n}(\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}_P^0})^{-1/2} \dot{\ell}_P^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \left[\mathbb{V}(\sqrt{n}N^{-1}\boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0} | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_P^0)\right]^{-1/2} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\eta}_i^{P\widehat{\boldsymbol{\beta}}_P^0} \to \mathbb{N}(\mathbf{0}, \mathbf{I}) \quad \text{(A.62)}
$$

in distribution.

Secondly, we exam the distance between $\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}_P^0}$ and $\mathbf{V}_P$. Here we introduce some new notations. Let $c_{\text{sub}}(i)$ denotes the rank of $t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)$ in $\{t_{(i)}^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)\}_{i=1}^N$, and $c_{\text{full}}(i)$ as the rank of $t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})$ in $\{t_{(i)}^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})\}_{i=1}^N$. We use g_{full} to represent the g in Theorem 3.1 and use g_{sub} to represent the quantity with same expression as g except replacing $\widehat{\boldsymbol{\beta}}_{\text{full}}$ with $\widehat{\boldsymbol{\beta}}_P^0$. Besides,

$$
\begin{aligned}
S_1 &= \{i | c_{\text{sub}}(i) \leq N - g_{\text{sub}} \,\&\, c_{\text{full}}(i) \leq N - g_{\text{full}}\}, \\
S_2 &= \{i | c_{\text{sub}}(i) \geq N - g_{\text{sub}} + 1 \,\&\, c_{\text{full}}(i) \leq N - g_{\text{full}}\}, \\
S_3 &= \{i | c_{\text{sub}}(i) \leq N - g_{\text{sub}} \,\&\, c_{\text{full}}(i) \geq N - g_{\text{full}} + 1\}, \\
S_4 &= \{i | c_{\text{sub}}(i) \geq N - g_{\text{sub}} + 1 \,\&\, c_{\text{full}}(i) \geq N - g_{\text{full}} + 1\}.
\end{aligned}
$$

We assume $S_2 \cup S_3 \cup S_4 = \emptyset$, then

$$
\begin{aligned}
&\left| \frac{1}{\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)} \right| \\
&= \left| \frac{\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)} \right| \\
&\leq \left| \frac{\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} \right| + \left| \frac{\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)} \right| \\
&\leq \frac{\sum_{j=1}^N \left| t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}}) - t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0) \right|}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} + \left| \frac{1}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)} \right| \sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0) \\
&= \Delta_1 + \Delta_2. \qquad \text{(A.63)}
\end{aligned}
$$

Note that combining with (A.37), we have

$$
\begin{aligned}
&\left|\|s_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\| - \|s_j(\widehat{\boldsymbol{\beta}}_P^0)\|\right| \\
=& \frac{\left|\|s_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\|^2 - \|s_j(\widehat{\boldsymbol{\beta}}_P^0)\|^2\right|}{\|s_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\| + \|s_j(\widehat{\boldsymbol{\beta}}_P^0)\|} \\
\leq& \frac{\sum_{k=1}^K \left|\left\{\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^2 - \left\{\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_P^0)\right\}^2\right|}{2K^{-\frac{1}{2}}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)^{-1}} \\
=& \frac{\sum_{k=1}^K \left|\left\{p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_{\text{full}}) - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_P^0)\right\}\left\{2\delta_{j.k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_{\text{full}}) - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_P^0)\right\}\right|}{2K^{-\frac{1}{2}}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)^{-1}} \\
\leq& \frac{2\sum_{k=1}^K \left|p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_{\text{full}}) - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_P^0)\right|}{K^{-\frac{1}{2}}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)^{-1}} \\
\leq& \frac{2\sum_{k=1}^K \|\dot{p}_k(\mathbf{x}_j, \boldsymbol{\varphi})\|\|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}_P^0\|\|\mathbf{x}_j\|}{K^{-\frac{1}{2}}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)^{-1}} \\
\leq& 2K^{3/2}\|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}_P^0\|\left(1 + Ke^{\lambda\|\mathbf{x}_j\|}\right)\|\mathbf{x}_j\|,
\end{aligned} \tag{A.64}
$$

where $\boldsymbol{\varphi} = u\widehat{\boldsymbol{\beta}}_{\text{full}} + (1-u)\widehat{\boldsymbol{\beta}}_P^0$, $u \in [0,1]$, $\dot{p}_k(\mathbf{x}_j, \boldsymbol{\beta})$ is the gradient of $p_k(\mathbf{x}_j, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and the last inequality is from the fact that $\|\dot{p}_k(\mathbf{x}_j, \boldsymbol{\beta})\| \leq 1$. Thus based on (A.37) and (A.64),

$$
\begin{aligned}
\Delta_1 =& \frac{\sum_{j=1}^N \left|t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}}) - t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)\right|}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} \\
\leq& \frac{1}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}\left\{2K^{3/2}\|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}_P^0\|\sum_{j=1}^N\left(1 + Ke^{\lambda\|\mathbf{x}_j\|}\right)\|\mathbf{x}_j\|^2\right\} \\
\leq& \frac{\sqrt{K}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)}{\|\mathbf{x}_i\|}\left\{2K^{3/2}\|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}_P^0\|\sum_{j=1}^N\left(1 + Ke^{\lambda\|\mathbf{x}_j\|}\right)\|\mathbf{x}_j\|^2\right\} \\
=& 2K^2\|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}_P^0\|\frac{\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)}{\|\mathbf{x}_i\|}\sum_{j=1}^N\left(1 + Ke^{\lambda\|\mathbf{x}_j\|}\right)\|\mathbf{x}_j\|^2
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta_2 =& \left|\frac{1}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)}\right|\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0) \\
=& \left|\frac{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0) - t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}\right|\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2K^{3/2}\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)\|\widehat{\boldsymbol{\beta}}_{\text{full}}-\widehat{\boldsymbol{\beta}}_P^0\|\|\mathbf{x}_i\|^2}{t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)t_i^{\text{optL}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)\\
&\leq \frac{2K^{3/2}\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)\|\widehat{\boldsymbol{\beta}}_{\text{full}}-\widehat{\boldsymbol{\beta}}_P^0\|\|\mathbf{x}_i\|^2}{\|\mathbf{x}_i\|^2}K\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)^2\sum_{j=1}^N t_j^{\text{optL}}(\widehat{\boldsymbol{\beta}}_P^0)\\
&\leq 2K^3\|\widehat{\boldsymbol{\beta}}_{\text{full}}-\widehat{\boldsymbol{\beta}}_P^0\|\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)^3\sum_{j=1}^N\|\mathbf{x}_j\|. \qquad \text{(A.65)}
\end{aligned}
$$

Thus combining with (A.63) and Lemma A.1.4, we have

$$
\begin{aligned}
\|\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}_P^0}-\mathbf{V}_P\| &= \frac{1}{N^2}\sum_{i=1}^N\|\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^T)\|\left|\frac{1-n\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})}-\frac{1-n\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}\right|\\
&= \frac{1}{N^2}\sum_{i=1}^N\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|^2\|\mathbf{x}_i\|^2\left|\frac{1}{\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})}-\frac{1}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}\right|\\
&\leq \frac{K}{N^2}\sum_{i=1}^N\|\mathbf{x}_i\|^2\left|\frac{1}{\pi_i(\widehat{\boldsymbol{\beta}}_{\text{full}})}-\frac{1}{\pi_i(\widehat{\boldsymbol{\beta}}_P^0)}\right|\\
&\leq \frac{K}{N^2}\sum_{i=1}^N\|\mathbf{x}_i\|^2\Big\{2K^2\|\widehat{\boldsymbol{\beta}}_{\text{full}}-\widehat{\boldsymbol{\beta}}_P^0\|\frac{\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)}{\|\mathbf{x}_i\|}\sum_{j=1}^N\left(1+Ke^{\lambda\|\mathbf{x}_j\|}\right)\|\mathbf{x}_j\|^2\\
&\quad +2K^3\|\widehat{\boldsymbol{\beta}}_{\text{full}}-\widehat{\boldsymbol{\beta}}_P^0\|\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)^3\sum_{j=1}^N\|\mathbf{x}_j\|\Big\}\\
&= \|\widehat{\boldsymbol{\beta}}_{\text{full}}-\widehat{\boldsymbol{\beta}}_P^0\|O_P(1)\\
&= O_{P|\mathcal{D}_N}(n_0^{-1/2}).
\end{aligned}
$$

Until now, we have proved that if $S_2\cup S_3\cup S_4=\emptyset$, $\|\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}_P^0}-\mathbf{V}_P\|\to 0$ in probability conditionally on $\mathcal{D}_N$. Next, we release the condition $S_2\cup S_3\cup S_4=\emptyset$. Note that under assumption 5 and Law of Large Numbers, we have

$$
\frac{1}{\frac{1}{N}\sum_{i=1}^N K^{-1/2}(1+Ke^{\lambda\|\mathbf{x}_i\|})^{-1}\|\mathbf{x}_i\|}-\frac{1}{\frac{1}{N}\sum_{i=1}^N\mathbb{E}\{K^{-1/2}(1+Ke^{\lambda\|\mathbf{x}_i\|})^{-1}\|\mathbf{x}_i\|\}}\to 0 \qquad \text{(A.66)}
$$

in probability. Noting that $n=o(N/\ln N)$ and combining with Lemma A.1.1, (A.37) and (A.66), for any $\boldsymbol{\beta}\in\Theta$, we have

$$
\begin{aligned}
\pi_{(N)}^o-\frac{1}{n} &= \frac{t_{(N)}^{\text{optL}}(\boldsymbol{\beta})}{\sum_{i=1}^N t_{(i)}^{\text{optL}}(\boldsymbol{\beta})}-\frac{1}{n}\\
&\leq \frac{\sqrt{K}\|\mathbf{x}\|_{(N)}/N}{\frac{1}{N}\sum_{i=1}^N K^{-1/2}(1+Ke^{\lambda\|\mathbf{x}_i\|})^{-1}\|\mathbf{x}_i\|}-\frac{1}{n}\\
&= O_P\left(\frac{\ln N}{N}\right)-\frac{1}{n}
\end{aligned}
$$

$$= \frac{1}{n}\left\{o_P(1) - 1\right\}, \tag{A.67}$$

where $\pi^o_{(N)}$ is the largest order statistic of $\{\pi^o_i\}_{i=1}^N$, which is defined in Theorem 3.1 except treating H as infinity and replacing $\widehat{\boldsymbol{\beta}}_{\text{full}}$ to $\boldsymbol{\beta}$. The result indicates that $\mathbb{P}\{\pi^o_{(N)} - \frac{1}{n} < 0\} \to 1$ and then we get $g \to 0$ in probability. Therefore, $\mathbb{P}(S_2 \cup S_3 \cup S_4 = \emptyset) \to 1$. Thus, we obtain

$$\|\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}^0_P} - \mathbf{V}_P\| = O_{P|\mathcal{D}_N}(n_0^{-1/2}). \tag{A.68}$$

Finally, from (A.43) and (A.59) in Lemma A.1.3, we have

$$\widehat{\boldsymbol{\beta}}^{ada}_P - \widehat{\boldsymbol{\beta}}_{\text{full}} = -\left\{\ddot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{-1}\dot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}}) + o_{P|\mathcal{D}_N}(1), \tag{A.69}$$

$$\left\{\ddot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{-1} - \mathbf{M}_N^{-1} = -\mathbf{M}_N^{-1}\left[\left\{\ddot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\} - \mathbf{M}_N\right]\left\{\ddot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{-1} = O_{P|\mathcal{D}_N}(n^{-1/2}) \tag{A.70}$$

and

$$\begin{aligned}
&\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\left(\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}^0_P}\right)^{1/2}\left[\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\left(\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}^0_P}\right)^{1/2}\right]^T \\
&= \mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}^0_P}\mathbf{M}_N^{-1}\mathbf{V}^{-1/2} \\
&= \mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_P\mathbf{M}_N^{-1}\mathbf{V}^{-1/2} + O_{P|\mathcal{D}_N}(n_0^{-1/2}) \\
&= \mathbf{I} + O_{P|\mathcal{D}_N}(n_0^{-1/2}).
\end{aligned} \tag{A.71}$$

Finally gathering (A.59), (A.62), (A.68), (A.69) and (A.70), by Slutsky's Theorem [Theorem 6 of 2], we have

$$\begin{aligned}
&\sqrt{n}\mathbf{V}^{-1/2}\left(\widehat{\boldsymbol{\beta}}^{ada}_P - \widehat{\boldsymbol{\beta}}_{\text{full}}\right) \\
&= -\mathbf{V}^{-1/2}\left\{\ddot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\}^{-1}\sqrt{n}\dot{\ell}^{*ada}_P + o_{P|\mathcal{D}_N}(1) \\
&= -\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\sqrt{n}\dot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}}) - \mathbf{V}^{-1/2}\left[\left\{\ddot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}})\right\} - \mathbf{M}_N^{-1}\right]\sqrt{n}\dot{\ell}^{*ada}_P + o_{P|\mathcal{D}_N}(1) \\
&= -\mathbf{V}^{-1/2}\mathbf{M}_N^{-1}\left(\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}^0_P}\right)^{1/2}\left(\mathbf{V}_c^{P\widehat{\boldsymbol{\beta}}^0_P}\right)^{-1/2}\sqrt{n}\dot{\ell}^{*ada}_P(\widehat{\boldsymbol{\beta}}_{\text{full}}) + o_{P|\mathcal{D}_N}(1) \\
&\to \mathbb{N}(\mathbf{0}, \mathbf{I}).
\end{aligned}$$

A.2 Asymptotic Properties of $\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}}$

Here, we present the asymptotic properties of $\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}}$, which is the final estimator of optimal subsampling with replacement algorithm in [4].

Theorem A.2.1 *Under Assumptions 1, 2 and 5, if $n_0/\sqrt{n} \to 0$, then as $n_0, n, N \to \infty$,*

$$\sqrt{n}\mathbf{V}_S^{-1/2}\left(\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\right) \to \mathbb{N}(\mathbf{0}, \mathbf{I}) \tag{A.72}$$

in distribution conditionally on $\mathcal{D}_N$ *and* $\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$*, in which* $\mathbf{V}_S = \mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}$ *with* $\mathbf{V}_{Nc}$ *having the expression of*

$$\mathbf{V}_{Nc} = \frac{1}{N^2}\left[\sum_{i=1}^{N}\frac{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})}{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_i\|}\right]\left[\sum_{j=1}^{N}\|\mathbf{s}_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_j\|\right].$$

Proof of Theorem A.2.1

First, we clarify some notations which are going to use in this proof. For optimal subsampling with replacement algorithm, the sample size is non-random. We use n_0 to denote the first stage sample size and n to denote the second stage sample size. The first stage subsample estimator is defined as $\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$ and the optimal subsampling probabilities under L-optimality are

$$\pi_{s,i}^{\text{optL}} = \frac{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_i\|}{\sum_{j=1}^{N}\|\mathbf{s}_j(\widehat{\boldsymbol{\beta}}_{\text{full}})\|\|\mathbf{x}_j\|}, i = 1, 2, ..., N. \tag{A.73}$$

Assume $n_0/\sqrt{n}\to 0$, the contribution of the first step subsample to the likelihood function is a small term with an order $o_{P|\mathcal{D}_N}(\sqrt{n})$ relative to the likelihood function. Thus, we can focus on the second step subsample only. Denote the log-likelihood function as

$$\ell_s^{*\text{ada}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N\pi^*_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}\left[\sum_{k=1}^{K}\delta^*_{i,k}\boldsymbol{\beta}_k^T\mathbf{x}_i^* - \log\left\{1+\sum_{l=1}^{K}e^{\boldsymbol{\beta}_l^T\mathbf{x}_i^*}\right\}\right], \tag{A.74}$$

where $\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})$ has the same expression as $\pi_{s,i}^{\text{optL}}$ except that $\widehat{\boldsymbol{\beta}}_{\text{full}}$ is replaced by $\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$. All quantities with * in (A.74) are from the second stage sample. For example, $\{\mathbf{x}_i^*\}_{i=1}^n$ mean the covariates of the second stage sample and $\{\pi^*_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}_{i=1}^n$ are the corresponding approximated optimal subsampling probabilities.

Before proofing, we need several lemmas.

Lemma A.2.2 *Under Assumptions 1 and 5, for* $k_2\geq 1$ *and* $k_1-k_2\geq -1$*,*

$$\frac{1}{N^{k_2+1}}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^{k_1}}{\pi^{k_2}_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}\leq K^{k_2}\left\{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^{k_1-k_2}\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)^{k_2}\right\}\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^{k_2}\right)=O_P(1), \tag{A.75}$$

where $\lambda=\sup_{\boldsymbol{\beta}\in\Theta}\|\boldsymbol{\beta}\|$.

Proof (**Lemma A.2.2**) Firstly, it is seen that

$$\begin{aligned}\left[\sum_{k=1}^{K}\{\delta_{i,k}-p_k(\mathbf{x}_i,\boldsymbol{\beta})\}^2\right]^{1/2} &\geq K^{-\frac{1}{2}}\sum_{k=1}^{K}|\delta_{i,k}-p_k(\mathbf{x}_i,\boldsymbol{\beta})|\\ &= K^{-\frac{1}{2}}\left\{\frac{1+\sum_{l=1}^{K}e^{\mathbf{x}_i^T\boldsymbol{\beta}_l}I(l\neq j)}{1+\sum_{k=1}^{K}e^{\mathbf{x}_i^T\boldsymbol{\beta}_k}}+\frac{\sum_{\substack{k=1\\k\neq j}}^{K}e^{\mathbf{x}_i^T\boldsymbol{\beta}_k}}{1+\sum_{k=1}^{K}e^{\mathbf{x}_i^T\boldsymbol{\beta}_k}}\right\}\end{aligned}$$

$$
\begin{aligned}
&= K^{-\frac{1}{2}} \left\{ \frac{1 + 2\sum_{l=1}^{K} e^{\mathbf{x}_i^T \boldsymbol{\beta}_l}(1 - \delta_{i,l})}{1 + \sum_{k=1}^{K} e^{\mathbf{x}_i^T \boldsymbol{\beta}_k}} \right\} \\
&\geq K^{-\frac{1}{2}} \frac{1}{1 + \sum_{k=1}^{K} e^{\mathbf{x}_i^T \boldsymbol{\beta}_k}} \\
&\geq K^{-\frac{1}{2}} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{-1}. \qquad \text{(A.76)}
\end{aligned}
$$

With Assumption 5 and Law of Large Numbers, we have

$$
\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{k_2} \\
&\leq \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(2Ke^{\lambda \|\mathbf{x}_i\|}\right)^{k_2} \\
&\leq (2K)^{k_2} \left(\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{2(k_1 - k_2)} \right)^{1/2} \left(\frac{1}{N} \sum_{i=1}^{N} e^{2k_2 \lambda \|\mathbf{x}_i\|} \right)^{1/2} \\
&= O_P(1), \qquad \text{(A.77)}
\end{aligned}
$$

where the last inequality is derived according to Cauchy-Schwarz inequality.

When $k_2 > 1$, combining with (A.76), Lemma 5 and Hölder inequality, we have

$$
\begin{aligned}
&\frac{1}{N^{k_2+1}} \sum_{i=1}^{N} \frac{\|\mathbf{x}_i\|^{k_1}}{\pi_{s,i}^{k_2}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^0)} \\
&= \frac{1}{N^{k_2+1}} \sum_{i=1}^{N} \frac{\|\mathbf{x}_i\|^{k_1 - k_2}}{[\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0)\}]^{k_2/2}} \left[\sum_{i=1}^{N} \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0)\}^2} \|\mathbf{x}_i\| \right]^{k_2} \\
&\leq \frac{K^{k_2/2}}{N^{k_2+1}} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{k_2} \left(\sum_{i=1}^{N} \left[\sum_{k=1}^{K} \left\{ \delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0) \right\}^2 \right]^{\alpha/2} \right)^{k_2/\alpha} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_2} \\
&\leq \frac{K^{k_2/2}}{N^{k_2+1}} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{k_2} \left(NK^{\alpha/2}\right)^{k_2/\alpha} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_2} \\
&= K^{k_2} \left\{ \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_1 - k_2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{k_2} \right\} \left(\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i\|^{k_2} \right) \\
&= O_P(1), \qquad \text{(A.78)}
\end{aligned}
$$

where $\alpha = \frac{k_2}{k_2 - 1}$.

When $k_2 = 1$,

$$
\frac{1}{N^2} \sum_{i=1}^{N} \frac{\|\mathbf{x}_i\|^{k_1}}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^0)}
$$

$$
\begin{aligned}
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^{k_1-1}}{[\sum_{k=1}^{K}\{\delta_{i,k}-p_k(\mathbf{x}_i,\widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}]^{1/2}}\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{K}\{\delta_{i,k}-p_k(\mathbf{x}_i,\widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2}\|\mathbf{x}_i\| \\
&\leq \frac{K}{N^2}\sum_{i=1}^{N}\|\mathbf{x}_i\|^{k_1-1}\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)\sum_{i=1}^{N}\|\mathbf{x}_i\| \\
&= \frac{K}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^{k_1-1}\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\| \\
&= O_P(1). \qquad \text{(A.79)}
\end{aligned}
$$

Finally, (A.75) is obtained due to (A.77), (A.78) and (A.79).

Lemma A.2.3 *If Assumptions 1, 2 and 5 hold, then*

$$\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} - \mathbf{M}_N = O_{P|\mathcal{D}_N}(n^{-1/2}) \qquad \text{(A.80)}$$

and

$$\dot{\ell}_s^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = O_{P|\mathcal{D}_N}(n^{-1/2}), \qquad \text{(A.81)}$$

where

$$\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} = \frac{\partial^2\ell_s^{*\text{ada}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{1}{nN}\sum_{i=1}^{n}\frac{\boldsymbol{\phi}_i^*(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i^*\mathbf{x}_i^{*\text{T}})}{\pi_{s,i}^*(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}.$$

Proof (**Lemma A.2.3**) Firstly, by directly calculation,

$$
\begin{aligned}
\mathbb{E}\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}|\mathcal{D}_N\right) &= \mathbb{E}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\left\{\mathbb{E}(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}^0_{\text{sub}})\right\} \\
&= \mathbb{E}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}(\mathbf{M}_N|\mathcal{D}_N) = \mathbf{M}_N, \qquad \text{(A.82)}
\end{aligned}
$$

where $\mathbb{E}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}$ means the expectation is taken with respect to the distribution of $\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$ given $\mathcal{D}_N$.

For any element $[\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}]^{j_1j_2}$ of $\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}$ where $1\leq j_1,j_2\leq dK$,

$$
\begin{aligned}
\mathbb{V}\left([\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}]^{j_1j_2}|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}^0_{\text{sub}}\right) &= \frac{1}{nN^2}\sum_{i=1}^{N}\frac{[\{\boldsymbol{\phi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\otimes(\mathbf{x}_i\mathbf{x}_i^{\text{T}})\}^{j_1j_2}]^2}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} - \frac{1}{n}(\mathbf{M}_N^{j_1j_2})^2 \\
&\leq \frac{1}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}, \qquad \text{(A.83)}
\end{aligned}
$$

where the last inequality holds by the fact that all elements of $\boldsymbol{\phi}_i$ are between 0 and 1.

From Lemma A.2.2 and (A.83),

$$\mathbb{V}\left([\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}]^{j_1j_2}|\mathcal{D}_N\right) = \mathbb{E}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\left\{\mathbb{V}\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}^0_{\text{sub}}\right)\right\}$$

$$+\mathbb{V}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\left\{\mathbb{E}\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}^0_{\text{sub}}\right)\right\} \tag{A.84}$$

$$\leq \mathbb{E}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\left\{\frac{1}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^4}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}\right\}$$

$$\leq \mathbb{E}_{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\left\{\frac{K}{n}\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)\right)\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|\right)\right\}$$

$$=\frac{K}{n}\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3\left(1+Ke^{\lambda\|\mathbf{x}_i\|}\right)\right)\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|\right)$$

$$=O_P(n^{-1}). \tag{A.85}$$

Using Markov's inequality, (A.80) follows from (A.82) and (A.85).

Similarly, we can achieve that

$$\mathbb{E}\left\{\frac{\partial\ell_s^{*\text{ada}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}}\Big|\mathcal{D}_N\right\}=0, \tag{A.86}$$

$$\mathbb{V}\left\{\frac{\partial\ell_s^{*\text{ada}}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial\boldsymbol{\beta}}\Big|\mathcal{D}_N\right\}=O_P(n^{-1}). \tag{A.87}$$

Finally, (A.81) is obtained combined with (A.86), (A.87) and Markov's inequality.

Lemma A.2.4 *If Assumptions 1, 2 and 5 hold, then*

$$\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}}-\widehat{\boldsymbol{\beta}}_{\text{full}}=O_{P|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}(n^{-1/2}) \tag{A.88}$$

Proof (**Lemma A.2.4**) Firstly, according to Lemma A.2.2, for any $\boldsymbol{\beta}\in\Theta$, we have

$$\begin{aligned}\mathbb{E}\left\{\ell_s^{*\text{ada}}(\boldsymbol{\beta})-\ell_f(\boldsymbol{\beta})\Big|\mathcal{D}_N,\widehat{\boldsymbol{\beta}}^0_{\text{sub}}\right\}^2&=\frac{1}{n}\left\{\frac{1}{N^2}\sum_{i=1}^{N}\frac{q_i^2(\boldsymbol{\beta})}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}-\ell_f^2(\boldsymbol{\beta})\right\}\\&\leq\frac{1}{n}\left\{\frac{1}{N^2}\sum_{i=1}^{N}\frac{2C_1^2\|\mathbf{x}_i\|^2+C_2^2}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}\right\}\\&\leq\frac{2C_1^2}{nN^2}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i\|^2}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}+\frac{2C_2^2}{nN^2}\sum_{i=1}^{N}\frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}\\&=O_P(1),\end{aligned} \tag{A.89}$$

where $C_1=\lambda(K+1)$, $C_2=1+\log K$, $\lambda=\sup_{\boldsymbol{\beta}\in\Theta}\|\boldsymbol{\beta}\|$ and

$$\begin{aligned}|q_i(\boldsymbol{\beta})|&=\left|\sum_{k=1}^{K}\delta_{i,k}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k-\log\left\{1+\sum_{l=1}^{K}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_l}\right\}\right|\\&\leq\sum_{k=1}^{K}\|\mathbf{x}_i\|\|\boldsymbol{\beta}_k\|+\log\left\{1+\sum_{l=1}^{K}e^{\|\mathbf{x}_i\|\|\boldsymbol{\beta}_l\|}\right\}\end{aligned}$$

$$
\begin{aligned}
&\leq K\|\mathbf{x}_i\|\|\boldsymbol{\beta}\| + \log\left(1 + Ke^{\|\mathbf{x}_i\|\|\boldsymbol{\beta}\|}\right) \\
&\leq K\|\mathbf{x}_i\|\|\boldsymbol{\beta}\| + 1 + \log K + \|\mathbf{x}_i\|\|\boldsymbol{\beta}\| \\
&\leq \lambda(K+1)\|\mathbf{x}_i\| + 1 + \log K \\
&= C_1\|\mathbf{x}_i\| + C_2.
\end{aligned}
$$

Therefore, from Lemma A.2.2 and (A.89),

$$
\mathbb{E}\left\{\ell_s^{*\text{ada}}(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta})|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0\right\}^2 = O_P(n^{-1}). \tag{A.90}
$$

Now, from (A.90), we have $\ell_s^{*\text{ada}}(\boldsymbol{\beta}) - \ell_f(\boldsymbol{\beta}) \to 0$ in conditional probability given $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}_{\text{sub}}^0$. Note that the parameter space is compact, and $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}$ and $\widehat{\boldsymbol{\beta}}_{\text{full}}$ are the global maximus of the continuous concave functions $\ell_s^{*\text{ada}}(\boldsymbol{\beta})$ and $\ell_f(\boldsymbol{\beta})$, respectively. Thus, conditionally on $\mathcal{D}_N$,

$$
\|\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\| = o_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0}(1), \tag{A.91}
$$

which ensures that $\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}$ is close to $\widehat{\boldsymbol{\beta}}_{\text{full}}$ as long as n is large enough.

Secondly, using Taylor's theorem (c.f. Chapter 4 of Ferguson 1996),

$$
0 = \dot{\ell}_{s,j}^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}}) = \dot{\ell}_{s,j}^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) + \frac{\partial \dot{\ell}_{s,j}^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}})}{\partial \boldsymbol{\beta}^T}(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}}) + R_j^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}, \tag{A.92}
$$

where $\dot{\ell}_{s,j}^{*ada}(\boldsymbol{\beta})$ is the partial derivative of $\ell_s^{*\text{ada}}(\boldsymbol{\beta})$ with respect to β_j, and

$$
R_j^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0} = (\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}})^T \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_{s,j}^{*ada}\{\widehat{\boldsymbol{\beta}}_{\text{full}} + uv(\widehat{\boldsymbol{\beta}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v \text{d}u\text{d}v \ (\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}}). \tag{A.93}
$$

The second derivative of $\dot{\ell}_{s,j}^{*ada}(\boldsymbol{\beta})$ satisfies the following condtion

$$
\left\| \frac{\partial^2 \dot{\ell}_{s,j}^{*ada}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\| \leq \frac{2}{n} \sum_{i=1}^n \frac{L\|\mathbf{x}_i^*\|^3}{N\pi_{s,i}^*(\widehat{\boldsymbol{\beta}}_{\text{sub}}^0)} = O_P(1),
$$

where L is a positive constant and the last equality is valid because, by Markov's Inequality and Assumption 5,

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n \frac{\|\mathbf{x}_i^*\|^3}{N\pi_{s,i}^*(\widehat{\boldsymbol{\beta}}_{\text{sub}}^0)} \geq \tau \middle| \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0\right) &\leq \frac{1}{nN\tau}\sum_{i=1}^n \mathbb{E}\left(\frac{\|\mathbf{x}_i^*\|^3}{\pi_{s,i}^*(\widehat{\boldsymbol{\beta}}_{\text{sub}}^0)} \middle| \mathcal{D}_N, \widehat{\boldsymbol{\beta}}_{\text{sub}}^0\right) \\
&= \frac{1}{N\tau}\sum_{i=1}^N \|\mathbf{x}_i\|^3 \to 0
\end{aligned} \tag{A.94}
$$

as $\tau \to \infty$. Thus from (A.93) we have

$$R_j^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} = O_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}} \left(\|\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\|^2 \right). \tag{A.95}$$

Finally, from (A.92) and (A.95),

$$\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}} = -\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} \right)^{-1} \left\{ \dot{\ell}_s^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}}(\|\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\|^2) \right\}. \tag{A.96}$$

From Lemma A.2.3, $\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} \right)^{-1} = O_{P|\mathcal{D}_N}(1)$. Combining this with (A.82), (A.91) and (A.96)

$$\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}} = O_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}}(n^{-1/2}) + o_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}}(\|\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\|),$$

which implies that

$$\widehat{\boldsymbol{\beta}}^{\text{ada}}_{\text{sub}} - \widehat{\boldsymbol{\beta}}_{\text{full}} = O_{P|\mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}}(n^{-1/2}). \tag{A.97}$$

Proof (**Theorem A.2.1**) Denote

$$\dot{\ell}_s^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{s}_i^*(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i^*}{N \pi_{s,i}^*(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}. \tag{A.98}$$

Given $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$, $\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} (i = 1, 2, ..., n)$ are independent random variables, with mean $\mathbf{0}$ and variance

$$\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} \equiv \mathbb{V}(\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}})}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} = O_P(1), \tag{A.99}$$

where the last equality holds because each element of $\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}$ is bounded by $N^{-2} \sum_{i=1}^{N} \pi(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})^{-1} \|\mathbf{x}_i\|^2$ and $N^{-2} \sum_{i=1}^{N} \pi(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})^{-1} \|\mathbf{x}_i\|^2$ is of order $O_P(1)$ from Lemma A.2.2. Meanwhile, for every $\varepsilon > 0$ and some $\rho > 0$,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\{\|n^{-1/2} \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\|^2 I(\|\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\| > n^{1/2}\varepsilon) | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}\} \\
&\leq \frac{1}{n^{1+\rho/2} \varepsilon^{\rho}} \sum_{i=1}^{n} \mathbb{E}\{\|\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\|^{2+\rho} I(\|\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\| > n^{1/2}\varepsilon) | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}\} \\
&\leq \frac{1}{n^{1+\rho/2} \varepsilon^{\rho}} \sum_{i=1}^{n} \mathbb{E}(\|\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}\|^{2+\rho} | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}) \\
&= \frac{1}{n^{\rho/2}} \frac{1}{N^{2+\rho}} \frac{1}{\varepsilon^{\rho}} \sum_{i=1}^{N} \frac{\|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|^{2+\rho} \|\mathbf{x}_i\|^{2+\rho}}{\pi_{s,i}^{1+\rho}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})}
\end{aligned}$$

$$\leq \frac{1}{n^{\rho/2}} \frac{1}{N^{2+\rho}} \frac{1}{\varepsilon^{\rho}} \sum_{i=1}^{N} \frac{K^{2+\rho}\|\mathbf{x}_i\|^{2+\rho}}{\pi_{s,i}^{1+\rho}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} = o_P(1),$$

where the last equality is from Lemma A.2.2. From (A.98) and (A.99), by the Lindeberg-Feller central limit theorem [Proposition 2.27 of 1],

$$\sqrt{n}(\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}})^{-1/2} \dot{\ell}_s^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) = \frac{1}{\sqrt{n}} \left[\mathbb{V}(\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} | \mathcal{D}_N, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\right]^{-1/2} \sum_{i=1}^{n} \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} \to \mathbb{N}(\mathbf{0}, \mathbf{I})$$

in distribution conditionally on $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$.

On the other hand, we exam the distance between $\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}}$ and $\mathbf{V}_{Nc}$. For clarity, we use $\{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{full}})\}_{i=1}^N$ to indicate $\{\pi_{s,i}^{\text{optL}}\}_{i=1}^N$ whose expression are shown in (A.73). First, it is seen that

$$\begin{aligned}
\|\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} - \mathbf{V}_{Nc}\| &= \frac{1}{N^2} \sum_{i=1}^{N} \|\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes (\mathbf{x}_i \mathbf{x}_i^T)\| \left| \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} \right| \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}}) \otimes \mathbf{x}_i\|^2 \left| \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} \right| \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \|\mathbf{s}_i(\widehat{\boldsymbol{\beta}}_{\text{full}})\|^2 \|\mathbf{x}_i\|^2 \left| \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} \right| \\
&\leq \frac{K^2}{N^2} \sum_{i=1}^{N} \|\mathbf{x}_i\|^2 \left| \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} \right|. \qquad \text{(A.100)}
\end{aligned}$$

For the last term in the above inequality,

$$\begin{aligned}
&\left| \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{1}{\pi_{s,i}(\widehat{\boldsymbol{\beta}}^0_{\text{sub}})} \right| \\
\leq &\left| \frac{\sum_{j=1}^N \sqrt{\sum_{k=1}^K \{\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} \|\mathbf{x}_j\|}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} \|\mathbf{x}_i\|} - \frac{\sum_{j=1}^N \sqrt{\sum_{k=1}^K \{\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2} \|\mathbf{x}_j\|}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} \|\mathbf{x}_i\|} \right| \\
&+ \left| \frac{\sum_{j=1}^N \sqrt{\sum_{k=1}^K \{\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2} \|\mathbf{x}_j\|}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} \|\mathbf{x}_i\|} - \frac{\sum_{j=1}^N \sqrt{\sum_{k=1}^K \{\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2} \|\mathbf{x}_j\|}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2} \|\mathbf{x}_i\|} \right| \\
\leq &\left| \frac{\sum_{j=1}^N \left(\sqrt{\sum_{k=1}^K [\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}_{\text{full}})]^2} - \sqrt{\sum_{k=1}^K [\delta_{j,k} - p_k(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})]^2} \right) \|\mathbf{x}_j\|}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} \|\mathbf{x}_i\|} \right| \\
&+ \frac{\sqrt{K} \sum_{j=1}^N \|\mathbf{x}_j\|}{\|\mathbf{x}_i\|} \left| \frac{1}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2}} - \frac{1}{\sqrt{\sum_{k=1}^K \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2}} \right| \qquad \text{(A.101)}
\end{aligned}$$

Note that, by (A.76),

$$
\begin{aligned}
&\left| \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} - \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2} \right| \\
&= \frac{\left| \sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2 - \sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2 \right|}{\sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} + \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2}} \\
&\leq \frac{\sum_{k=1}^{K} \left| \left\{ \delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}}) \right\}^2 - \left\{ \delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}) \right\}^2 \right|}{2K^{-\frac{1}{2}} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{-1}} \\
&= \frac{\sum_{k=1}^{K} \left| \left\{ p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}}) - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}) \right\} \left\{ 2\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}}) - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{P}) \right\} \right|}{2K^{-\frac{1}{2}} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{-1}} \\
&\leq \frac{2 \sum_{k=1}^{K} \left| p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}}) - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}}) \right|}{K^{-\frac{1}{2}} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{-1}} \\
&\leq \frac{2 \sum_{k=1}^{K} \|\dot{p}_k(\mathbf{x}_j, \boldsymbol{\varphi})\| \|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}^0_{\text{sub}}\| \|\mathbf{x}_i\|}{K^{-\frac{1}{2}} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^{-1}} \\
&\leq 2K^{3/2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right) \|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}^0_{\text{sub}}\| \|\mathbf{x}_i\|, \qquad \text{(A.102)}
\end{aligned}
$$

where $\boldsymbol{\varphi} = u\widehat{\boldsymbol{\beta}}_{\text{full}} + (1-u)\widehat{\boldsymbol{\beta}}^0_{\text{sub}}$, $u \in [0,1]$, and $\dot{p}_k(\mathbf{x}_j, \boldsymbol{\beta})$ is the gradient of $p_k(\mathbf{x}_j, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The last inequality is from the fact that $\|\dot{p}_k(\mathbf{x}_j, \boldsymbol{\beta})\| \leq 1$. Combning (A.76) and (A.102), we have

$$
\begin{aligned}
&\left| \frac{1}{\sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2}} - \frac{1}{\sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2}} \right| \\
&= \frac{\left| \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} - \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2} \right|}{\sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})\}^2} \sqrt{\sum_{k=1}^{K} \{\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^0_{\text{sub}})\}^2}} \\
&\leq 2K^{3/2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right) \left\| \widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}^0_{\text{sub}} \right\| \|\mathbf{x}_i\| \sqrt{K} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right) \sqrt{K} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right) \\
&\leq 2K^{5/2} \left(1 + Ke^{\lambda \|\mathbf{x}_i\|}\right)^3 \left\| \widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}^0_{\text{sub}} \right\| \|\mathbf{x}_i\|. \qquad \text{(A.103)}
\end{aligned}
$$

Thus, from (A.100), (A.101), (A.102) and (A.103),

$$
\|\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}^0_{\text{sub}}} - \mathbf{V}_{Nc}\| \leq \|\widehat{\boldsymbol{\beta}}_{\text{full}} - \widehat{\boldsymbol{\beta}}^0_{\text{sub}}\| C = O_{P|\mathcal{D}_N}(n_0^{-1/2}), \qquad \text{(A.104)}
$$

where

$$
\begin{aligned}
C = 2K^{5/2}\Bigg\{ & \frac{1}{N}\sum_{i=1}^{N} \frac{\|\mathbf{x}_i\|}{\sqrt{\sum_{k=1}^{K}[\delta_{i,k} - p_k(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\text{full}})]}} \cdot \frac{1}{N}\sum_{i=1}^{N}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)\|\mathbf{x}_i\|^2 \\
& + K^{3/2}\frac{1}{N}\sum_{i=1}^{N}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)^3\|\mathbf{x}_i\|^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|\Bigg\} \\
\leq 2K^3\Bigg\{ & \frac{1}{N}\sum_{i=1}^{N}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)\|\mathbf{x}_i\| \cdot \frac{1}{N}\sum_{i=1}^{N}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)\|\mathbf{x}_i\|^2 \\
& + \frac{K}{N}\sum_{i=1}^{N}\left(1 + Ke^{\lambda\|\mathbf{x}_i\|}\right)^3\|\mathbf{x}_i\|^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|\Bigg\} = O_P(1),
\end{aligned}
$$

and the last equality follows from (A.77) and Lemma 5.

Finally, from (A.96), (A.97) and Lemma A.2.3,

$$
\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}} = -\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{-1}\dot{\ell}_s^{*ada}(\widehat{\boldsymbol{\beta}}_{\text{full}}) + O_{P|\mathcal{D}_N}(n^{-1}), \tag{A.105}
$$

$$
\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{-1} - \mathbf{M}_N^{-1} = -\mathbf{M}_N^{-1}\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}_{\text{sub}}^0} - \mathbf{M}_N\right)\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{-1} = O_{P|\mathcal{D}_N}(n^{-1/2}) \tag{A.106}
$$

and

$$
\begin{aligned}
& \mathbf{V}_S^{-1/2}\mathbf{M}_N^{-1}\left(\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{1/2}\left[\mathbf{V}_S^{-1/2}\mathbf{M}_N^{-1}\left(\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{1/2}\right]^{\mathrm{T}} \\
& = \mathbf{V}_S^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\mathbf{M}_N^{-1}\mathbf{V}_S^{-1/2} \\
& = \mathbf{V}_S^{-1/2}\mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}\mathbf{V}_S^{-1/2} + O_{P|\mathcal{D}_N}(n_0^{-1/2}) \\
& = \mathbf{I} + O_{P|\mathcal{D}_N}(n_0^{-1/2}).
\end{aligned} \tag{A.107}
$$

Further, based on Lemma A.2.3, (A.104), (A.105), (A.106), (A.107) and Slutsky's Theorem [Theorem 6 of 2], we achieve

$$
\begin{aligned}
& \sqrt{n}\mathbf{V}_S^{-1/2}\left(\widehat{\boldsymbol{\beta}}_{\text{sub}}^{\text{ada}} - \widehat{\boldsymbol{\beta}}_{\text{full}}\right) \\
& = -\mathbf{V}_S^{-1/2}\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{-1}\sqrt{n}\dot{\ell}_s^{*ada} + O_{P|\mathcal{D}_N}(n^{-1/2}) \\
& = -\mathbf{V}_S^{-1/2}\mathbf{M}_N^{-1}\sqrt{n}\dot{\ell}_s^{*ada} - \mathbf{V}_S^{-1/2}\left\{\left(\mathbf{M}_n^{*\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{-1} - \mathbf{M}_N^{-1}\right\}\sqrt{n}\dot{\ell}_s^{*ada} + O_{P|\mathcal{D}_N}(n^{-1/2}) \\
& = -\mathbf{V}_S^{-1/2}\mathbf{M}_N^{-1}\left(\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{1/2}\left(\mathbf{V}_c^{\widehat{\boldsymbol{\beta}}_{\text{sub}}^0}\right)^{-1/2}\sqrt{n}\dot{\ell}_s^{*ada} + O_{P|\mathcal{D}_N}(n^{-1/2}) \\
& \to \mathbb{N}(\mathbf{0}, \mathbf{I})
\end{aligned}
$$

in distribution conditionally on $\mathcal{D}_N$ and $\widehat{\boldsymbol{\beta}}_{\text{sub}}^0$.

References

[1] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, London, 1998.

[2] Thomas S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall, Los Angeles, USA, 1996.

[3] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 1999.

[4] Yaqiong Yao and HaiYing Wang. Optimal subsampling for softmax regression. *Statistical Papers*, 60(2):585–599, 2019.